Combining Multiple Scoring Systems for Target Tracking Using Rank-Score Characteristics

Damian M. Lyons and D. Frank Hsu

Robotics & Computer Vision Laboratory
Department of Computer & Information Science
Fordham University
Bronx NY 10458
{lyons, hsu}@cis.fordham.edu

Keywords: Sensor fusion, Target tracking, Video analysis, Combinatorial Fusion Analysis (CFA), Multiple scoring systems, Rank-Score function, Rank combination, Score combination, Feature selection.

Abstract

Video target tracking is the process of estimating the current state, and predicting the future state of a target from a sequence of video sensor measurements. Multitarget video tracking is complicated by the fact that targets can occlude one another, affecting video feature measurements in a highly non-linear and difficult to model fashion. In this paper, we apply a multisensory fusion approach to the problem of multitarget video tracking with occlusion. The approach is based on a data-driven method (CFA) to selecting the features and fusion operations that improve a performance criterion.

Each sensory cue is treated as a scoring system. Scoring behavior is characterized by a rank-score function. A diversity measure, based on the variation in rank-score functions, is used to dynamically select the scoring systems and fusion operations that produce the best tracking performance. The relationship between the diversity measure and the tracking accuracy of two fusion operations, a linear score combination and an average rank combination, is evaluated on a set of twelve video sequences. These results demonstrate that using the rank-score characteristic as a diversity measure is an effective method to dynamically select scoring systems and fusion operations that improve the performance of multitarget video tracking with occlusions.

1. Introduction

Automated tracking of targets in video has a number of applications, including automated surveillance, robotics and virtual reality, amongst others. However, it remains a difficult problem, especially when handling video with multiple targets and crowded scenes [13]. For example, a video camera looking at an airport lobby or a busy city intersection will have exactly this kind of scene, and this motivates our interest in finding an approach to tracking that works well in such cases.

A video image can be a very rich source of information about a target: image position, image velocity, color properties, shape properties and so forth. Fusing these multiple sources of information is an appealing way to make tracking more robust [41]. Existing approaches to sensory fusion for video tracking have tended to fall into one of three categories: statistical approaches, physical modeling approaches and heuristic approaches. In this paper based on our previous work ([15]-[17][24]), we propose a new approach using Combinatorial Fusion Analysis (CFA) (Hsu, Chung and Kristal [14]) which has been applied to other fields such as information retrieval, pattern recognition, virtual screening and drug discovery, and protein structure prediction (see for example, [12] [17]-[18] [22] [26] [28] [43]). This approach is bottom-up and data-driven. It develops methods and criteria for dynamically selecting feature subsets and fusion operations that improve a performance measure. Because this approach does not make assumptions about what targets can and cannot do, it can be applied successfully to situations

such as video tracking with multiple mutual target occlusions where it is difficult to formulate a computationally efficient physical or statistical model.

Section 2 is a review of related literature. Section 3 presents our framework for combining multiple scoring systems. In Section 4, we motivate the importance of this approach for tracking with multiple mutual target occlusions in the process of target hypothesis pruning and feature selection. Experimental results are reported in Section 5. Twelve video sequences containing a variety of tracking situations form the basis of the experiments. Section 6 presents conclusions and Section 7 discusses future plans.

2. Related Work

Previous work in fusion for multisensory video tracking can be divided into three categories [25]: statistical, physical and heuristic. The first, and arguably the largest, category represents sensory measurements as random variables whose probability density functions can be characterized and used to define a sensory fusion operation. The target tracking community has developed a number of such elegant approaches [1]. The Kalman filter is an example of one of the earliest developed approaches, where the sensor measurement noise is a random variable characterized by a zero-mean Gaussian distribution. Reid's Multiple Hypothesis Tracking (MHT) algorithm [33] extends this approach to handle multitarget point tracking (e.g., radar targets), and Cox and Hingorani [5] reported an efficient implementation of this for tracking video corner features. However, video tracking rarely meets assumptions of Gaussian zero-mean noise. Sharma [37] developed a general Bayesian framework for fusion, presenting Maximum Likelihood (ML) and Maximum A Posteriori (MAP) formulations. In general, in a Bayesian approach, it is assumed that the different feature measurements are conditionally independent, and therefore that the conditional probability of an estimated quantity S given a collection of image data I can be expressed using Bayes rule. In the standard framework for linear estimation, this gives rise to an estimate for S that is a linear combination of the cue measurements where the combination coefficients are inversely proportional to the variance.

Rasmussen and Hager [32] build on another target-tracking algorithm, the Joint Probability Density Association Filter (JPDAF) which uses multiple visual cues to track multi-part objects. They develop a Joint Likelihood Filter (JLF), an extension of JPDAF to the multisensory case, where a joint likelihood is a product of component feature likelihoods in conjunction with a relative depth mask. The depth mask addresses the problem of target occlusion and its non-linear effect on feature measurements. Borghys et al. [3] use logistic regression to find parameters for the conditional probabilities of a target given a set of (texture) feature measurements. These parameters are then used as weights in a linear combination to yield a fused feature estimate.

The second category of work considers that if the image generation process can be modeled in sufficient detail, then this physical model can be used to determine how sensory measurements should be fused. Nandhakumar and Aggarwal [25] use a physics-based modeling approach to develop fusion formula for infrared, optical and sonar measurements. The feature measurements are combined, often in a non-linear fashion, to produce meaningful physical measurements that can be used to recognize objects.

The final category of work is the heuristic category. The fusion of data in this case is based on a proposed heuristic measurement, derived from a pragmatic appreciation of the nature of the problem. Checka and Wilson [4] adopt an approach to the fusion of audio and video information for tracking in which the video information is used to coarsely localize the person, and the audio information is then used to derive a more accurate localization. Loy et al. [23] use a particle filter approach to represent multiple target hypotheses. To fuse their multiple visual cues, they employ a weighted sum of cue measurements, where each cue is weighted by a reliability coefficient that measures how closely the PDF of each cue has approximated the fused PDF. (The reliability is also used to allocate computational resources across cues.) Snidaro et al. [38] use an appearance ratio (AR) to determine the reliability of a sensor. The AR value is used to weight the position estimates from a sensor. Triesch and von der Malsburgh [39] again define fusion as a weighted sum of local cue measurement, where each cue estimate

is weighted by a reliability coefficient. The dynamics of the reliability coefficients are phrased generally, and lead to a majority consensus style of fusion.

The statistical and physical modeling approaches both rely on being able to correctly and efficiently model the relationship between feature measurements and target state. However, when one or more targets engage in repeated mutual occlusions, the relationship between targets and feature measurements can become highly non-linear and very difficult to model. The heuristic approaches sidestep this problem by adopting an approximate, rather than exact statistical or physical model. The disadvantage is, however, that there is no guarantee of performance.

An alternate approach is to estimate a statistical or physical model from a collection of sensor and associated ground truth measurements. Rao [30][31] assumes that the output of each sensor is related to the actual feature values by an unknown probability distribution. A sample of independent and identically distributed pairs of actual feature values and sensor outputs for each value is collected, and Rao addresses how to use this information to select a fusion rule from a collection that performs within a specified bound of the best fusion rule from the collection. He shows that this bound is related to the length of the sample and develops a polynomial time algorithm to estimate a best fusion rule.

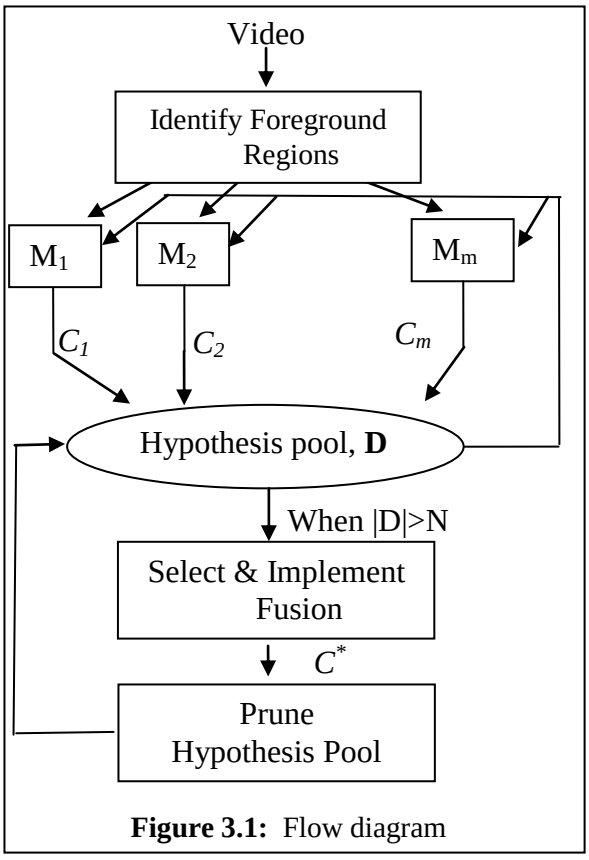


Figure 3.1: Flow diagram

More recently in [15][17][24], the authors have proposed and studied a dynamic and efficient approach to fusion for multitarget tracking in CCTV surveillance (called RAF – Rank and Fuse). Experimental results were obtained to illustrate the use of the RAF approach, explaining the advantage of rank versus score combinations of features for each target.

The work in this paper uses the framework of combining multiple scoring systems, Combinatorial Fusion Analysis (CFA), and the rank-score function (Hsu, Chung and Kristal [14] and Hsu and Taksa [19]) as a measure of diversity between scoring systems. This is different from previous statistical or physical modeling approaches. It has several distinct characteristics that distinguish it from existing fusion methods (see e.g., [6] [41]-[42]). It is bottom-up and does not impose a model on the measurements. The multiple scoring systems represent different sensory cues, features, or combinations of cues and/or features. In this paper, we consider: (a) both score and rank function for each feature or piece of evidence to be combined, and explore the interaction between the two functions by calculating the rank-score function, and (b) the rank-score function as a measure of diversity between scoring characteristics to select a best fusion operation. We use ground-truth information to validate the approach.

3. Combination of Multiple Scoring Systems

We have proposed a multiple hypothesis framework for implementing and evaluating a variety of feature fusion operations, including score and rank fusion combinations, for video tracking applications [17] and [24]. In this paper, we enhance and update that framework and use it as a basis for our experimentation. The framework is shown in Figs. 3.1 and 3.2. Video information is preprocessed to extract foreground regions and then channeled to one or more tracking modules, $M_1,...,M_m$. Each module,

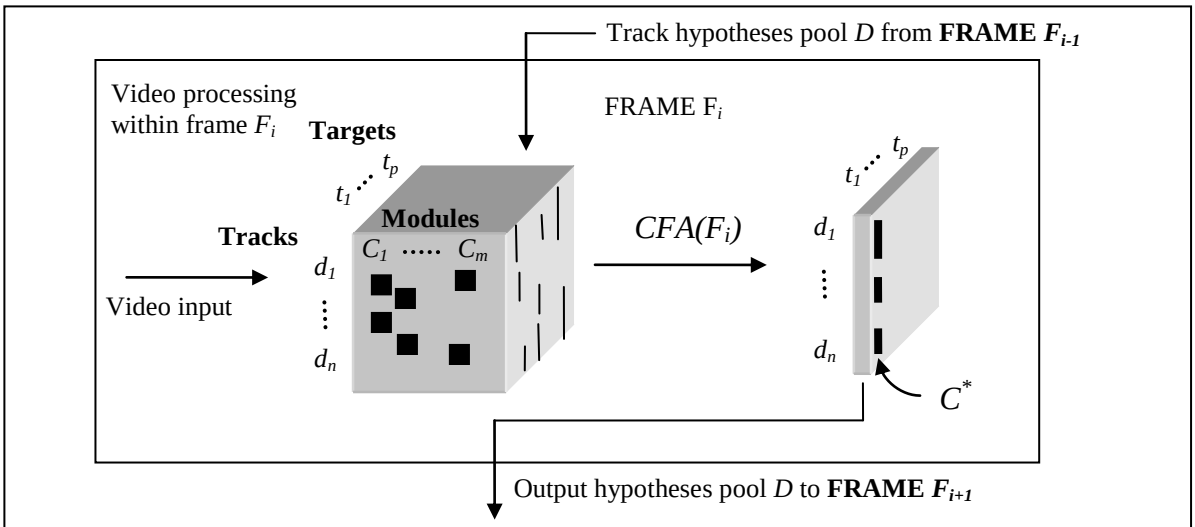


Figure 3.2: Combination of multiple scoring systems where d_h, $h \in [1, n]$ are track hypotheses; C_k, $k \in [1, m]$ and t_j, $j \in [1, p]$, are scoring module k and target j respectively.

M_k, uses the video information to produce a set of track lists for each target j, C_{kj}. The modules may employ different sensory cues, features, combinations of features and/or different association approaches to produce the target list. We consider each tracker as a *scoring system* on the set of target track hypotheses.

Each target list C_k is a list of the hypothesized tracks, for each of the $1,\ldots, p$ targets, produced by module M_j given the previous set of track hypotheses and the current video segmentation, $C_k = \bigcup_{j=1}^{p} C_{kj}$.

The number of targets, p, may vary as tracking proceeds, and the multiple hypothesis tacking approach handles target track initiation and termination. Each track hypothesis includes a score value that captures how well the track matches the target from the perspective of the information and approach used by the tracking module. The hypothesis pool $\boldsymbol{D}$ is the union of all the track lists for each target, a set of track hypotheses for each target labeled with scores from each of the tracking modules, $\boldsymbol{D} = \bigcup_{k=1}^{m} C_k$.

The data space of module scores and track hypothesis for each target, for each video frame, is shown in Fig. 3.2. The hypothesis pool is allowed to grow until it reaches a threshold size N, at which point fusion is performed. The fusion operation is shown in the target data space cube shown in Fig. 3.2. The selection and implementation of the fusion operations in $CFA(F_i)$ will be dealt with in more detail below. The fused target list C^* is then pruned to the q best ranked candidate tracks for each target, and the lower ranked candidates are discarded.

3.1 Score and Rank Functions of a Scoring System Module.

For two integers, a and b where $a \le b$, we write $[a, b]$ for the set of all integers x, $a \le x \le b$. Let $\boldsymbol{D}_j = \{d_1,\ldots,d_n\} \subseteq \mathbf{D}$ be the track hypotheses in the pool of n track hypotheses for target $j \in [1, p]$ generated by the collection of scoring systems. We will assume that each module operates on the same pool of track hypotheses. This could be by use of a common hypothesis generation stage [24], as in our case, or by the generation of a set of composite tracks [27].

The score function $s_{kj}(d)$ assigns a real number to each d in $\boldsymbol{D}_j$ which is the score given by the tracking module M_k to the candidates for target j. When treating $s_{kj}(d)$ as an array of real numbers, it

would lead to a rank function $r_{kj}(d)$ after sorting the $s_{kj}(d)$ array into descending order and assigning a rank (a positive natural number) to each of the d in $\boldsymbol{D}_j$. The resulting rank function $r_{kj}(d)$is a function from $\boldsymbol{D}_j$ to $\boldsymbol{N}=[1, n]$ (we note that $|\boldsymbol{D}_j|=n$). There is a monotonic relationship:

$$[\, s_{kj}(d_1) > s_{kj}(d_2)\,] \Rightarrow [\, r_{kj}(d_1) < r_{kj}(d_2)\,]$$

There is ambiguity when two track hypotheses have the same score. To resolve this, we add the constraint:

$$[\, (s_{kj}(d_{i1}) = s_{kj}(d_{i2})) \wedge (i1 < i2)\,] \Rightarrow [\, r_{kj}(d_{i1}) < r_{kj}(d_{i2})]$$

Fig. 3.2 shows the feature fusion selection and implementation process in more detail. The target track list from module $k \in [1, m]$ for target $j \in [1, p]$, containing both rank and score information, is written C_{kj}. The feature selection process determines which subset of the m target track lists to use in fusion for target j. The fusion selection process determines which fusion operation to use to combine the selected features for target j. The output of the fusion framework is a fused target track list C_j^* containing the top q candidates for each target.

Normalization is needed to properly compare and correctly combine score functions from multiple scoring systems. We adopt the following transformation from $s_{kj}(d):\boldsymbol{D}\rightarrow\boldsymbol{R}$ to $s^*_{kj}(d):\boldsymbol{D}\rightarrow[0,1]$ where

$$s^*_{kj}(d) = \frac{s_{kj}(d) - s_{\min}}{s_{\max} - s_{\min}},\ d \in \boldsymbol{D} \text{ and } s_{max}=max\{\, s_{kj}(d)|\ d \in \boldsymbol{D}\} \text{ and } s_{min}=min\{\, s_{kj}(d)|\ d \in \boldsymbol{D}\}.$$

3.2. Rank and Score Combinations

Given m scoring systems for a target j with score functions $s_{kj}(d)$ and rank functions $r_{kj}(d)$ and $k \in [1, m]$, there exist several different ways of combining the output of the scoring systems, including score combination, rank combination, voting, average combination and weighted combination. For the m scoring systems with $s_{kj}(d)$ and $r_{kj}(d)$, we define the score functions s_R and s_S of the rank combination (RC) and score combination (SC) respectively as:

$$s_R(d) = \sum_{k=1}^{m}\left[w_k r_{kj}(d)\right], \text{ and } \quad s_S(d) = \sum_{k=1}^{m}\left[v_{kj} s_{kj}(d)\right].$$

As we did before, $s_R(d)$ and $s_S(d)$ are sorted into ascending and descending order to obtain the rank function of the rank combination $r_R(d)$ and the score combination $r_S(d)$, respectively. For this paper, we will define $w_k = 1/m$ and $v_k = \dfrac{1/\sigma_{kj}^2}{\sum_1^m 1/\sigma_{kj}^2}$ where σ_{kj}^2 is the variance of s_{kj}. That is, the rank combination is an average rank combination, and the score combination is a *Mahalanobis* combination.

We will adopt these two fusion rules as examples of *linear combination* and of *rank combination* rules. These two classes have been discussed widely in the literature to understand relative strengths and weaknesses: For example, Kittler and Alkoot [20] characterizes when a Vote combination outperforms a Sum in terms of the estimation error. Melnik, Vardi and Zhang [26] studies several rank-based combinations in a unifying framework. We adopt a Mahalanobis combination, rather than a general linear sum (weighted average), because of its relationship to the Bayesian formulation and its widespread use in tracking. We select average rank combination as a representative of rank-based rules such as voting, max, min etc. Rao [30] defines a *metafuser* as a fusion rule that combines the complementary performance of two kinds of fusion rules to produce a better performing fusion rule. Our objective will be to develop a rule to select for each target and video frame which of these two fusions will best improve tracking performance.

When m scoring systems, $k \in [1, m]$, together with the score functions $s_{kj}(d)$ and rank functions

$r_{kj}(d)$ are used, combinatorially there are 2^m-1 ($= \sum_{k=1}^{m} \binom{m}{k}$) possible combinations for these *m* scoring systems using either rank or score functions. The order of complexity is exponential and becomes prohibitive when *m* is large. The study of multiple scoring systems on large data sets ***D*** involves sophisticated mathematical, statistical, and computational approaches and techniques (see [14] and refs).

3.3. The Rank-Score Graph of a Scoring System Module

Hsu and Taksa [19] characterize the relationship that an expert habitually produces between score and rank as the *rank-score function*s and the graph of that function as the *rank-score graph* (Fig. 3.3)*;* the graph of a monotonic function *f* that relates the rank and score of a set of candidates. Let $s : \boldsymbol{D} \rightarrow \boldsymbol{R}$, where *s(d)* is the score of candidate *d* in the set of candidates ***D***. Let $r : \boldsymbol{D} \rightarrow \boldsymbol{N}$, where *r(d)* is the rank of candidate *d* when the candidates are ordered according to their score. Then, *the rank-score function f* is the composite of *s* and *r* defined as $f : \boldsymbol{N} \rightarrow \boldsymbol{R}$, where

$$f(i) = (s \circ r^{-1})(i) = s(r^{-1}(i)).$$

A rank-score graph has to be *monotonic non-increasing.* However, the shape of the graph can be different for different experts and is a characteristic of that expert's approach. An expert who assigns scores in a linearly decreasing fashion will have a linear rank-score graph (e.g., Fig. 3.3 (f_2)). An expert who habitually assigns high scores to a large subset of its top ranked candidates will have a graph that is not a straight line, but has a low slope for the top ranked candidates and a higher slope for the remainder. The concave-down graph f_3 in Fig. 3.3 is an example of this. A third kind of scoring behavior is exemplified by f_1 in Fig 3.3. In this case, the expert habitually gives higher scores to a small subset of its top ranked candidates and much lower scores to the rest.

Hsu and Taksa [19] indicate that a diversity measure based on the rank-score graph can be used to

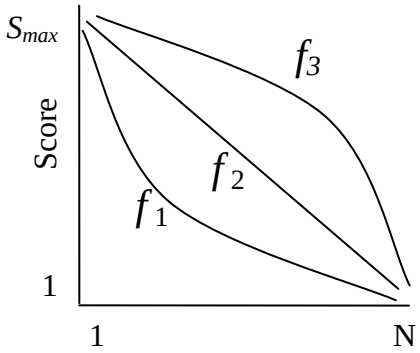


Figure 3.3: Three rank-score graphs

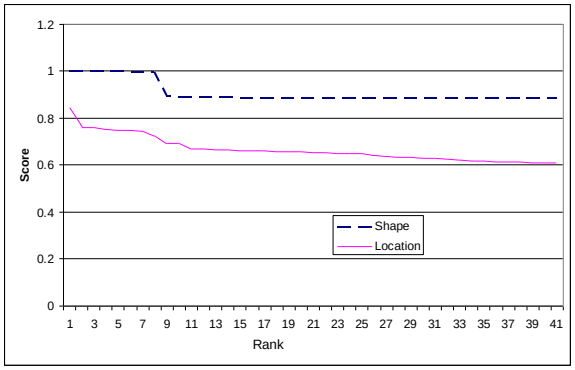


Figure 3.4: Example rank-score graphs generated by the RAF tracker. *Shape* refers to the shape video feature described in Section 5.1 and *Location* refers to the Location video feature. The score is the inverse of the similarity value defined in Section 5.1.

determine whether a score or rank fusion will produce a better result (see also [21] for other diversity measures). Hsu and colleagues ([15][19][22][43]) have used the new paradigm for diversity measurements between two scoring systems in a variety of applications. When the rank-score graphs of two experts are very similar, then a score combination will produce the best fusion. When the rank-score graphs are very different, then a rank combination produces the better result.

3.4. Diversity between Scoring System Module Characteristics.

Returning to the rank and score function definitions of Section 3.1, it is now possible to define a set of tracker rank-score functions. The rank score function for tracker module *k* for target *j* is:

$$f_{kj} : \mathbf{N} \rightarrow \mathbf{R},\ f_{kj}(i) = s_{kj}(r_{kj}^{-1}(i)) = \text{score of track hypothesis } d \in D_j \text{ which has rank } i$$

The rank-score graph of the scoring system module *k* for target *j* is the graph of the rank-score function f_{kj}. In the case of video tracking, the scoring behavior that is captured by the rank-score graph is a characteristic of the scoring system, which includes choice of cue or feature measurements, the video scene and the algorithm used by the tracker module. Fig. 3.4 shows two rank-score graphs from one tracking sequence (Sequence 9; other examples are presented in [24]). This example illustrates that rank-score graphs can have a variety of forms, based on the feature and/or fusion operators used as well as on the tracking scenario. Hsu and Taksa's results indicate that when the graphs for a target become sufficiently different, a rank fusion operation will most likely perform better than a score fusion operation.

We compare the rank-score graphs from each scoring system module for each target to determine which to use, and which fusion operation to employ. We define the difference between two rank score graphs f_A and f_B as follows (over N ranks):

$$d(f_A, f_B) = \sum_{i=1}^{N} (f_A(i) - f_B(i))$$

For two modules *A* and *B*, when $d(f_A, f_B)$ is sufficiently large, then we propose rank fusion will outperform score fusion for these two modules *A* and *B*. In Section 5 we evaluate this proposition experimentally by looking at the combinatorial combinations of the fusion operations and evaluating the relationship between this diversity measure and a ground-truth based performance measurement. The results of this study will demonstrate that this diversity measure is a useful criterion for selecting fusion operations.

4. Target Hypothesis Pruning and Feature Selection

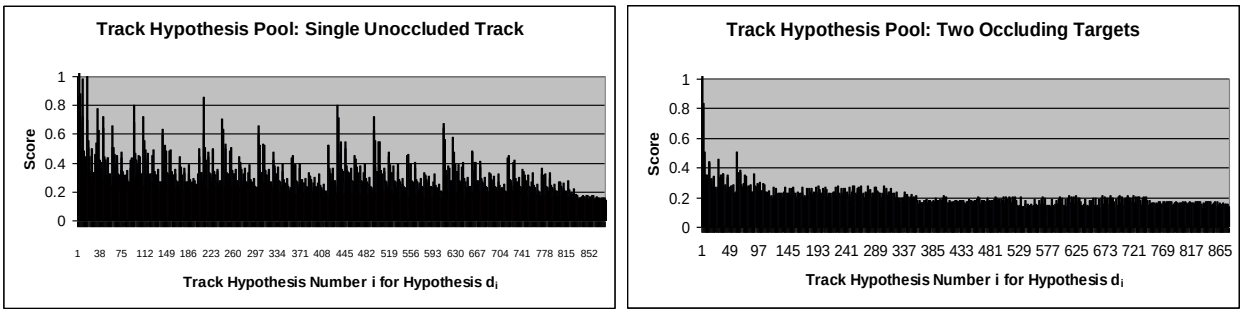
In this section, we describe the importance of, and rationale for using the rank-score function in the combination of multiple scoring systems for tracking in a scenario with repeated mutual target occlusion. In particular we compare this heavy occlusion scenario with a much simpler, unoccluded tracking scenario for two tasks important for feature combination in tracking: (a) target hypothesis pruning, and (b) feature selection. We show that in the heavy occlusion scenario, using rank and score combination has distinct advantages in target hypothesis pruning. On the other hand, we also show that the rank-score function and the variation of the rank-score function among individual scoring systems can be used to select features that improve the rate of false positives (FP) and false negatives (FN) of the combined scoring system.

4.1 Target Hypothesis Pruning

Hsu et al. [17] introduces the Rank-and-Fuse (RAF) multiple hypothesis video tracking framework as a way to investigate the combinatorial options for feature fusion. Experimental results are reported for three video sequences of a single target that splits into two separate but very similarly colored targets (see Lyons et al. [24]). A variety of fusion operators including fusion using rank as well as score combinations are evaluated on each sequence. The top 20 track hypotheses for each sequence are compared to ground truth. The best results, measured as the most correct tracks in the top 20 track hypotheses, are obtained by using the position feature and no fusion. However, the best fusion results are obtained with rank fusion operators, which out-perform the score fusion operator for this tracking scenario. A second experiment is conducted, adding a shape feature (target bounding box ratio), and inspecting the rank-score graphs for each feature. The shape rank-score graph is of a different overall form than that of position or color. They then verify Hsu and Taksa's conclusion [19] that a rank combination of shape and color outperforms a score combination in this situation.

Hsu and Lyons [16] explore some of the theoretical implications of rank versus score in tracking. Figure 4.1 shows examples of typical track hypothesis score distributions for two tracking scenarios. The graph data were collected with the RAF tracker using the position tracking feature module (tracking the location of the centroid of each foreground region). The track hypothesis pool was logged in each case

after tracking had proceeded for approximately 15 frames. Tracking a single, unoccluded target produces the distribution in Fig. 4.1(a) (Sequence 3 in Section 5). The distribution in Fig. 4.1(b) is the result of tracking two targets that engaged in repeated mutual occlusions; two people walking as a couple (Sequence 7 in Section 5). The single target tracking scenario produces a greater variance in scores, because the scoring system can distinguish good target hypotheses. There is less variance observed for the crowded tracking scenario because the scoring system has difficulty distinguishing good and bad hypotheses; the correct choice of target is much less clear cut.



(a) Single, unoccluded target (b) Two partially occluding targets

Figure 4.1: Typical score distributions for two tracking scenarios

In [16] we are motivated by this to propose that the track hypothesis scoring distributions are different in these cases, and to derive the rank-score graph associated with the scoring system for each of the two scenarios. This analysis, presented in Appendix A, illustrates that for a crowded scene the benefit of score based fusions and of rank-based fusions will vary depending on the hypothesis pool pruning threshold. This explains why in crowded tracking scenarios, working with rank and score combinations has a distinct advantage, because the same score cutoff produces more variations in rank in a scenario such as Fig. 4.1(a) than in one such as Fig. 4.1(b). So in that case, working directly with rank combinations can produce a more accurate result.

4.2 Feature Selection

The analysis of the previous section can be continued to understand the implication of rank versus score in selecting features for fusion when tracking in a crowded scenario by restricting our attention to scenarios such as Fig 4.1(b) but now considering *more* than one scoring system. In [16] we show that if scoring systems with *complementary* rank-score functions (e.g., Fig. 3.3 f_1 and f_3) are combined, they produce a better performing combination. We have used the number of false positives (FP) and false negatives (FN) associated with the combination as our criterion for evaluating performance. If scoring systems with complementary rank-score functions are combined, they produce a combination with a rank-score function which will minimize the false positives (FP) and false negatives (FN) associated with the combination. Appendix B presents a revised form of this argument.

Trackers with complementary rank-score graphs should be distinguished from trackers whose output is negatively correlated or independent. The latter is a relationship between the scores (i.e., the score function *s(d)* for *d* in *D*, the set of all track hypotheses) the trackers assign to a specific track. However, the former is a relationship between scoring behaviors (i.e., the rank-score function *f(i)* for *i* in $N=[1, n]$ and $|D|=n$), irrespective of the track being scored. Trackers may be correlated, negatively correlated or independent and still have complementary rank-score graphs. This gives the rank-score characteristic approach a distinctive advantage of characterizing the scoring behavior difference. It leads to a new approach to the quantitative and qualitative study of using, for example, the rank-score characteristics as diversity among multiple scoring systems.

5. Experiments

In this section, we describe the implementation of the video tracker (Section 5.1), and present two types of experimental results:

(1) **Type I Experiments (**Section 5.2.**):** These show that for crowded scenes, a mix of score combination fusions and rank combination fusions can produce a significantly better tracking result. The experiments do not say how to choose operators to produce the improvement.

(2) **Type II Experiments(**Section 5.3**):** These are the same as Type I except that we incorporate the rank-score graph information for selecting between fusions. They demonstrate that the difference of rank-score graphs criterion is an effective way to select which fusion operation to perform.

Sequence	Description	Frames
1	1 moving target, indoors	53
2	2 slowly crossing targets, indoors	40
3	1 moving target, outdoors	30
4	3 moving targets, outdoors, non-adjacent	23
5	5 targets in loose group, outdoors	40
6	4 moving targets, outdoors, 2 overlapping	20
7	2 targets moving as a couple, outdoors	104
8	7 targets moving as a crowd, outdoors	50

Table 5.1: Description of video sequences used in Type I experiments

We obtained ground truth information for twelve video sequences showing a variety of unrehearsed targets moving in one indoor (a lab) and one outdoor (a campus footpath) scene. The targets are not always easily separated from the background or each other, and in most sequences, they are close enough to each other to cause recurrent partial occlusions. Tables 5.1 and 5.4 describe each of the sequences in terms of whether they were indoor or outdoor, single or multiple targets, moving targets as a couple or as a crowd, and targets as a loose, overlapping or crossing group. However, in each sequence, some targets can be separated most of the time, unlike the dense crowds studied in [34]. Fig. 5.1 shows example frames from these sequences. Ground truth was obtained by having a human observer go through the video sequence frame by frame and annotate the position of each target. The twelve video sequences are available on the Fordham Robotics and Computer Vision Lab web site at http://www.cis.fordham.edu/rcvlab.

(a) (b)

Figure 5.1: (a) Four example frames showing Lab and Footpath backgrounds and varieties of targets from Sequences 2, 5, 7 and 8.
(b) Examples of mutual target occlusions from these sequences (showing extracted foreground regions only, with bounding boxes)

5.1 Implementation

The video tracking algorithm has four stages:

(1) Background subtraction: Generate potential target measurements by eliminating the background elements of the scene.

(2) Track hypothesis generation: Using the existing pool of hypothesis in conjunction with the new target measurements to generate a larger pool of hypotheses.

(3) Score generation: Score the pool of hypotheses using color, location and shape measurements.

(4) Fusion and pruning: Fuse multiple score information and reduce the track hypotheses pool size to include only the best hypotheses.

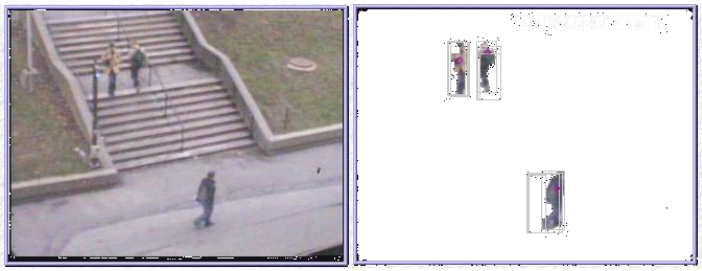


Figure 5.2: Example background subtraction.

Background Subtraction. Foreground regions are extracted from each frame of the image sequence using the non-parametric background estimation technique of Elgammal et al. [8]. A non-parametric distribution is learned for each pixel based on 5 to 10 seconds of video of an initial empty (but not static) scene. Background learning is suspended during target tracking, and pixels are classified as background or foreground based on where they fall with respect to the background distribution learned for that pixel. This approach is effective in filtering background phenomena that result in multimodal pixel value distributions including moving foliage, rain, a small amount of camera vibration and lighting changes.

Figure 5.2 shows an example of background subtraction for a three target outdoor sequence. Pixels classified as foreground are clustered using a connected components algorithm. Components above a threshold size are considered potential target regions. This is indicated in Fig. 5.1 with a gray bounding box; the multiple bounding boxes in the vicinity of each component indicate its position and shape in the previous four frames.

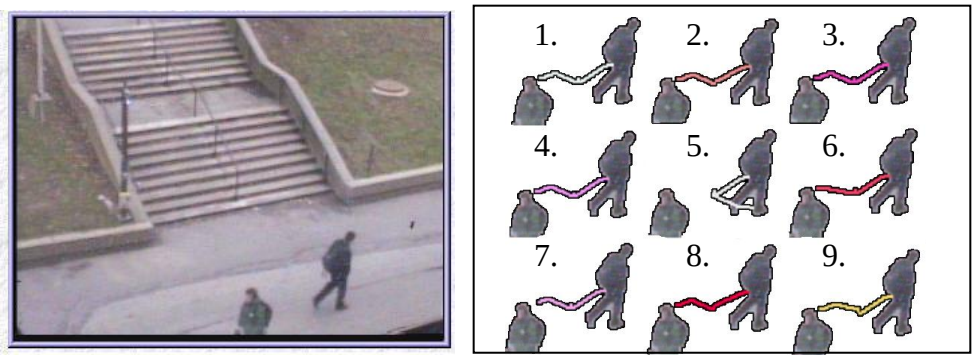


Figure 5.3: Multiple track hypothesis for two targets crossing; last frame in sequence (left), top 9 track hypotheses (rank order is left to right, top to bottom) drawn superimposed on foreground regions of last frame (right). Most of the top 9 (except rank 5) are similar, differing in start and end locations other smaller details.

Track hypotheses generation. Foreground regions are potential target measurements. For each frame *i* in the video sequence, a common MHT based [1][5] hypothesis generation module associates these measurements with the set of existing track hypothesis $\mathbf{D}_i$ to produce a new pool of all track hypotheses

(e.g., Fig 5.3). The gating function is that the position of the next component in a track hypothesis, $p(c_j)$ be within a standard deviation of the predicted position p_k for target k:

$$(p_k - p(c_j))^2 < \sigma_k^2$$

Any existing track hypothesis which meets the gating criterion for a component c_j is associated with that region, and a new track hypothesis is generated that is the old track extended by this region. In addition to the extension of tracks by new measurements, each region also gives rise to a new track of length 1 initialized to a fixed new track score (to model newly appearing targets), and each track gives rise to a new track of the same length with its score modified by a fixed false alarm score (to model false alarms). The pool of track hypotheses grows as follows:

$$| \boldsymbol{D}_{i+1} | = | \boldsymbol{D}_i | \times (n_i + 1) + n_i$$

where n_i is the number of regions segmented from frame i.

Score Generation. All three component trackers in the RAF system share the pool of track hypothesis. Each tracker traverses the pool and annotates each hypothesis with a score based on the features measured by that component tracker. For example, the color tracker stores an average normalized RGB value $(\bar{r}, \bar{g}, \bar{Y})$ for each track hypothesis, defined as

$$\bar{r} = \frac{1}{N_C}\sum_C r, \quad \bar{g} = \frac{1}{N_C}\sum_C g, \quad \bar{Y} = \frac{1}{N_C}\sum_C Y$$

where C is the image region of the target, N_c is the number of pixels in C, and r, g and Y are the normalized RGB values of a pixel in the image region. This value is compared to the average normalized RGB (r_j, g_j, Y_j) measured on a foreground component c_j using:

$$s_{col} = (r_j - \bar{r})^2 + (g_j - \bar{g})^2 + (Y_j - \bar{Y})^2$$

This similarity value is averaged over all the components (one per frame) of a track hypothesis to obtain the color score for that track hypothesis. Note that the similarity value is *smaller* for better hypotheses. Because this is the opposite convention to that usually adopted in CFA, all the rank-score graphs in this paper have been plotted used the inverse of the similarity score for consistency and clarity.

The scores for shape and position are calculated in a similar way:

- The shape measurement is the area covered by the target, and the shape similarity measure is the ratio of target area to foreground component area: $s_{sha} = N_C \big/ N_{C_j}$
- The location measurement is the image coordinates of the location of the centroid of the target region, and the location similarity is the Cartesian distance between target centroid and component centroid: $s_{loc} = | p(C) - p(c_j) |$

The set of target to measurement association hypotheses (including new targets and false alarms, and assuming that at most one measurements matches at most one target) is then generated and used to calculate a normalized score value for each track hypothesis.

Fusion and Pruning. The pool of track hypotheses grows combinatorially, and needs to be pruned to stay within resource limits. The resource limits are represented by a nominal pool size n_T:

$$(| \mathbf{D}_i | > k_T n_T) \Rightarrow \textit{Prune } \mathbf{D}_i \textit{ down to size } n_T$$

The values $n_T=100$, $k_T=2.5$ were used here. The top scoring candidates for all targets after fusion were preserved. To get the best track hypotheses for each target candidate set, the scores from each of the separate trackers are fused in two ways.

(a) Mahalanobis score fusion (*MS*): Let $s_{k,l}$ be the score for t_k by tracker l and $\sigma^2_{k,l}$ be the variance:

$$s_{k,bs} = (q_{k,col}\, s_{k,col} + q_{k,loc}\, s_{k,loc} + q_{k,sha}\, s_{k,sha}) \text{ where } q_{k,l} = \frac{\left(\frac{1}{\sigma^2_{k,l}}\right)}{\left(\frac{1}{\sigma^2_{k,col}}\right)+\left(\frac{1}{\sigma^2_{k,loc}}\right)+\left(\frac{1}{\sigma^2_{k,sha}}\right)}$$

(b) Average rank fusion (*AR*): Let $r_{k,l}$ be the rank of track hypothesis t_k according to tracker l :

$$s_{k,ar} = \frac{1}{3}\,(r_{k,co} + r_{k,loc} + r_{k,sha})$$

In each case, the top m=30 tracks produced by tracker were evaluated against ground truth using a Mean Sum of Squared Distances (MSSD):

$$\frac{1}{nm}\sum_j\sum_i (gp_i - tp_{ij})^2$$

where gp_i, $i \in [1, n]$ is the ground truth sequence of target centroid image locations and tp_{ij}, $i \in [1, n]$, is the *jth* best track's sequence of target centroid image locations. Whichever fusion scores *lower* by this measure is considered the better fusion and this is the one adopted for this target. If both score the same, then the score fusion was used. Different fusions may be adopted for different targets, and of course, a track hypothesis might have several different fusions used on it over the course of successive pruning events. The image sequence index number and type of fusion used is recorded for each track hypothesis.

Once the fusion calculation is completed, the top scoring track hypotheses for each target are kept, the rest are deleted, and the tracking continues.

5.2 Mixed Combinations for Type I Experiments

In the first experiment, the RAF tracking system was modified to carry out two fusion operations, a score fusion and an average rank fusion, both described in more detail below. The tracker was run *three times* on each of the eight video sequences; we will refer to these as RUN1, RUN2 and RUN3:

- In RUN1, single features were used for tracking; i.e., no fusion was performed. RUN1-A used the position feature only; RUN1-B, the color feature only; and, RUN1-C, shape feature only.
- In RUN2, a score fusion of all three features was carried out.
- In RUN3, the tracker was allowed to evaluate both rank and score fusion of all three features whenever a fusion needed to be performed, and selected between them as described below.

For each target, for each fusion, the top scoring 30 track hypotheses are evaluated against the ground truth data using the MSSD measure. In addition, the top 30 tracks were examined to see which fusion operators had been used. In RUN3, whenever a fusion needed to be performed, the fusion operator that produced *the better MSSD value on its top 30 tracks* at that point in the tracking process was selected.

Results

The combined (over all targets) MSSD average and variance for the single feature only runs (RUN1-A, -B, -C), and the score fusion of all three features run (RUN2), for each video sequence, are shown in Table 5.2. The MSSD performance of the score fusion of all three features (RUN2 as in the last column of table 5.2) and the mixed score and rank fusion run (RUN3) is shown in Table 5.3. The *lower* the MSSD value shown in the table, the closer the tracking results were to ground truth. In 5 of the 8 sequences (shaded row sequences, 1, 3-5, 7 in Table 5.2), use of a fusion operator (RUN2 or RUN3) is an improvement on the single feature tracking (RUN1). In all cases, the combination of score and rank fusion (RUN3) is as good as the score fusion only (RUN2).

Figure 5.4: Top scoring track for the right target in sequence 7 superimposed on the ground truth track for the target with three single features and the score fusion:
(a) Position-only, (b) color-only, (c) shape-only, and (d) score fusion of all three.

Fig. 5.4 shows the *best* track for the right-most target at the end of each of the four runs for sequence 7. While the numbers in Table 5.2 were calculated based on the top 30 tracks (not just the best single track, as shown here) for both targets, Fig. 5.4 illustrates typical errors in the single feature tracks vs. the fused track. Note that the targets walked as a single group until close to the center of the image; all features report very similar tracks for this section. After this, the position feature follows the ground truth track, but loses the target frequently. (Missing sections of track indicate that target was lost and then reacquired.) Color and shape produce erratic tracks and shape loses target quite frequently. The fused result is appears similar to the position result, but without as many target losses.

Seq.	**RUN1-A** Position Feature Only		**RUN1-B** Color Feature Only		**RUN1-C** Shape Feature Only		**RUN2** Score Fusion of all	
	MSSD Av.	**MSSD Var**	**MSSD Av.**	**MSSD Var**	**MSSD Av.**	**MSSD Var**	**MSSD Av.**	**MSSD Var**
1	1540.14	720.08	1564.88	1012.44	1538.19	764	1537.22	694.7
2	708.4	3306.29	726.21	3534.48	583.95	2453.85	816.53	8732.13
3	117.49	73.09	113.43	67.82	112.77	88.86	108.89	61.61
4	32.53	9.4	33.67	9.37	33.11	9.36	23.14	2.39
5	355.29	158.46	345.05	165.2	346.65	155.82	334.138	120.111
6	74.06	16.7	336.81	960.58	76.2	18.1	96.4	119.22
7	607.3	266.09	592.61	227.14	612.27	266.51	577.78	201.29
8	390.16	445.98	548.26	726.49	538.76	738.24	538.35	605.84

Table 5.2: MSSD results for single features and score fusion of all features.
RUN1-A, -B, -C show tracking performance using position, color and shape respectively. RUN2 shows tracking performance using a score fusion combination of all three features. Lower MSSD implies better tracking performance.

In 4 of the 8 sequences (the shaded row sequences, 2, 6-8 in Table 5.3) use of the combination of both score or rank fusion is a significant improvement over the use of score fusion only. Figure 5.5 illustrates the situation for the right hand target in sequence 7. The top 30 tracks for the three single feature cases and the two fused cases are shown overlaid on the ground truth track. The initial part of the tracks in every case are similar, since the targets move together until close to the center of the image. Many of the tracks for each case are similar (as in Fig 5.3) and almost completely overlay. As in Fig. 5.5, the score fusion result, RUN2, is similar to the position result except with fewer target losses. However, displaying the top 30 tracks for position and for score fusion shows that for the last quarter of the image, they consist of a mix of tracks for the right and for the left target. On the other hand, the ground-truth guided fusion, RUN3, has a collection of tracks that (not surprisingly) closer to the ground-truth right hand track.

Seq.	RUN2 Score fusion of all		RUN3 Score and rank fusion using ground truth to select		t value
	MSSD Avg.	**MSSD Var.**	**MSSD Avg.**	**MSSD Var.**	
1	1537.22	694.7	1536.65	695.49	0.1
2	816.53	8732.13	723.13	3512.19	4.65
3	108.89	61.61	108.34	60.58	0.23
4	23.14	2.39	23.04	2.30	0.186
5	334.138	120.111	332.89	119.39	0.44
6	96.4	119.22	66.9	12.91	8.12
7	577.78	201.29	548.6	127.78	15.5
8	538.35	605.84	500.9	57.91	9.08

Table 5.3: MSSD results for score fusion and ground truth selected mix of score and rank fusion. RUN2 shows tracking performance using a score fusion combination of all three features (as in Fig 5.2). RUN3 shows tracking performance when using ground truth to select between score and rank fusion. RUN3 results that improve on RUN2 with a significance > 95% are shaded.

Figure 5.5: Comparison of top 30 tracks overlaid for three single feature cases and two fused cases: (a) Position-only, (b) color-only, (c) shape-only, (d) score fusion of all 3 (RUN2), and (e) ground-truth selected score or rank fusion of all 3 (RUN3).

Of course it is possible that the difference in MSSD measurements for the fusions was due to chance. To address this, we calculate the t-test statistic [29] for the RUN2 and RUN3 fusion distributions. The results of the significance test are shown in the last column of Table 5.3. Results with a significance level of 95% or greater are exactly those shaded in row sequences 2, and 6-8.

Discussion

Although the MSSD for the combined score and rank fusion case (RUN3) is smaller for all 8 sequences, this difference is only significant at the 95% level or greater in four sequences, sequences 2, 6, 7, and 8. Looking at the description of these sequences in Table 5.1, they all share the characteristic that they include multiple, overlapping targets. We do not see a performance improvement in single targets (e.g., sequences 1 and 3) or tracking multiple targets (e.g., sequences 4 and 5). However, when there are multiple targets that move in such a way as to cause repeated partial occlusion, then we observe a

significant improvement from the additional use of a rank-based fusion operation. In this situation, there is a lot of splitting and merging of the image regions associated with the targets. In the video sequences without such effects, we produced no significant quality improvement. This phenomenon is consistent with our previous experiments and analytic work as described in Section 4.

This experiment demonstrates that including rank fusion in addition to score fusion can be valuable in tracking. However, this approach of including rank and score fusion alone could not be used as an algorithmic basis for a better tracker, as it requires the existence of ground truth data to select whether rank fusion or score fusion should be used. We need to develop a way to decide when to use rank fusion and when to use score fusion that is not based on knowledge of the ground truth. In the next section, Section 5.2, we will compute the rank-score function f_A of a tracker A and use the variation between f_A and f_B to guide us in the process of combination using rank fusion and score fusion.

5.3 Selection of Combination using the Rank-Score Characteristics for Type II Experiments

In Section 3.4 we have defined the rank-score diversity $d(f_A, f_B)$ of two rank-score functions f_A and f_B for trackers A and B. In our implementation we have three features. Let f_t be the rank-score graph for tracker t. We use the largest absolute difference between any two of the three features for selecting fusions:

$$\delta_{rs} = MAX \mid d(f_{t1}, f_{t2}) \mid for\ t1 \neq t2$$

Note that the rank-score graph for each target for each feature is computed dynamically from hypothesis score information. The null hypothesis in our type II experiments is that this maximum difference of rank-score graphs is the same for fusion events where the score fusion produced the better results as for fusion events where the rank fusion produced the better result. If we disprove the null hypothesis, then this maximum difference is a useful criterion for selecting between fusion operations.

Sequence	Description	Length
9	2 slowly crossing targets, outdoors,	28
10	2 adjacent moving targets, outdoors	43
11	9 targets in group, outdoors	35
12	1 quickly moving target, outdoors	34

Table 5.4: Description of video sequences added for Type II experiments

Seq.	RUN2 Score fusion		RUN3 Score and rank fusion using ground truth to select		RUN4 Score and rank fusion using rank-score function to select	
	MSSD Avg.	MSSD Var.	MSSD Avg.	MSSD Var.	MSSD Avg.	MSSD Var.
1	1537.22	694.47	1536.65	695.49	1536.9	694.24
2	816.53	8732.13	723.13	3512.19	723.09	3511.41
3	108.89	61.61	108.34	60.58	108.89	61.61
4	23.14	2.39	23.04	2.30	23.14	2.39
5	334.13	120.11	332.89	119.39	334.138	120.11
6	96.40	119.22	66.9	12.91	67.28	13.38
7	577.78	201.29	548.6	127.78	577.78	201.29
8	538.35	605.84	500.9	57.91	534.3	602.85
9	143.04	339.73	140.18	297.07	142.33	294.94
10	260.24	86.65	252.17	84.99	258.64	85.94
11	520.13	2991.17	440.98	2544.69	470.27	2791.62
12	1188.81	745.01	1188.81	745.01	1188.81	745.01

Table 5.5: MSSD results for Type II experiments
Lower MSSD implies better tracking performance.

In the final phase of this experiment, we identify a threshold value for the maximum difference measurement and we rerun the eight video sequences through the tracker but now using the maximum difference measurement (rather than the ground truth measurements) to select fusion operation. In addition, we run the tracker on 4 additional video sequences that were not used in the selection of the threshold operation. We compare these MSSD results with those from the first experiment.

Results

The ground-truth guided combination of score fusion and rank fusion (RUN3) of the type I experiments) was repeated, and the average and variance of the maximum difference of rank-score graphs was calculated separately for score and rank fusions for the four video sequences for which RUN3 showed a significant improvement (sequence 2, and sequences 6-8). The average value of the difference for the score fusion operator, $\delta_{rs} = 0.05$, was then selected as a threshold value for this second set of experiments. If the variation between the rank-score graphs is less than or equal to δ_{rs} then a score fusion is used, otherwise a rank fusion is applied.

All 8 original sequences and the 4 new sequences were run, and the MSSD performance figures collected. The results are shown in Table 5.5 labeled as RUN4. For convenience of comparison, the RUN2 and RUN3 figures from Table 5.3 are shown again in Table 5.5. Appendix C shows four typical rank-score graphs generated during RUN4.

Discussion

Table 5.5 shows that in all of the 12 sequences, the use of the variation between rank-score functions to select a fusion operator (RUN4) performed as good as the use of score fusion (RUN2). However, in 4 of the 12 sequences (2, 6, 8, and 11), RUN4 performed better than RUN2. Our conclusion from this is that the maximum variation between rank-score graphs is a useful predictor for which fusion operator to use to produce the best tracking performance. On the other hand, when we compare the rank-score selected fusion (RUN4) to the ground-truth selected fusion (RUN3) we see that the rank-score selected fusion performance does not always achieve the level of performance as the ground-truth selected fusion. This is not surprising as the ground-truth selected fusion has the advantage of knowing the correct target track before making its choice. However, in a real-time application, ground truth will not normally be known.

As such, our experiments confirm what Hsu and Taksa [19] proposed, that the rank-score function is a feasible and useful characteristic to guide us in the process of rank and score fusion. As can be seen from Table 5.5, RUN4 using the rank-score function selected fusion, and without knowing the ground-truth, performed as good as RUN3 except in three cases (sequences 7, 8, and 11). Even in these three cases, RUN4 is as good as RUN2 for sequences 7 and 8 and much better than RUN2 for sequence 11.

6. Summary and Conclusions

This paper presents a data-driven, combinatorial fusion analysis approach to the problem of selecting a multisensory fusion operation to improve the performance of multitarget video tracking with occlusion. Our tracking framework considers each feature measurement to be a separate *scoring system* on the set of target track hypotheses, and scoring behavior was characterized by the rank-score function. Two fusion operations were considered: an average rank fusion and a Mahalanobis score fusion. We proposed a measure of diversity $d(f_A, f_B)$ between two scoring systems (cues, features or tracking systems) A and B which is equal to the sum of differences between the two rank-score functions $f_A(i)$ and $f_B(i)$ across all ranking orders i in N. The measure of performance we used was the mean sum of squared differences (MSSD) between a hypothesized track and the ground-truth for the track, as established by a human observer.

We used 12 video sequences covering a variety of situations. Our results are summarized as follows:

(1) Combination using simple score fusion (RUN2) improved the performance over single feature tracking.

(2) Using ground-truth information to select a mix of rank and score fusion operations (RUN3) produces a significant improvement over score fusion.

(3) Using rank-score diversity to select a mix of rank and score fusion operations (RUN4) produces a better resultant than score fusion alone (RUN2) but not as good as the ground-truth selection (RUN3). However, since ground truth is not typically known, the rank-score diversity approach is more powerful.

More generally speaking, the CFA approach has several advantages. Among them:

(1) **Efficiency:** For each target, we use *n* tracks and *m* scoring systems, converting score functions to rank functions using fast sorting algorithms would require a maximum of *n*m*log n* steps. Rank fusion or score fusion using the average operation requires *n* additions only. As an example, on a 1.4GHz Pentium M laptop with 376 Mbytes of RAM running Windows™ 2000, RUN4 operated between 5 and 12 fps. This is somewhat conservative timing since a great deal of diagnostic and logging information was also being generated during the run.

(2) **Scalability:** Our method can be applied to the case of multiple scoring systems with a large number, *n*, of tracks. The number of scoring systems, *m*, can also be large. In that case, a subset of scoring systems would have to be selected to perform fusion.

(3) **Adaptiveness:** The fusion operation is dynamically selected from the set of fusion rules, to best suit the target and scene characteristics, as tracking proceeds.

(4) **Diversity:** Measurement of diversity between the multiple scoring systems (cues, features and systems) is explored to guide us in the selection of rank fusion or score fusion. In this paper, we use the variation between the rank-score functions.

(5) **Visualization:** The rank-score graph is a highly visual representation of scoring behavior and of diversity among multiple scoring systems.

7. Future Work

Our current paper represents one of the first in a series of on-going projects using the framework of CFA and the concept of a rank-score function in the study of target tracking and recognition, and the design of a robust, real-time and on-line intelligent system for such applications. Our study suggests several issues and directions for future work. These include:

(1) **Performance evaluation:** The MSSD measurement is used in this paper to evaluate the performance of a scoring system. In general, given two scoring systems, *A* and *B*, we like to find a criterion (or criteria) to predict the improvement of the combined scoring system *C(A,B)*. In this regards, the combination *C(A,B)* is seen as a positive case if the performance of *C*, *P(C)*, is greater than or equal to the performance of *A* and *B* (i.e., $P(C) \geq max\{P(A),P(B)\}$). Otherwise it is a negative case (see [28][43]). We have started a study along these lines ([15]).

(2) **Diversity measurement:** The difference of the rank-score functions f_A and f_B of two scoring systems *A* and *B* was used in this paper to represent the scoring diversity between *A* and *B*. That is, $d(A, B) = d(f_A, f_B)$. We will explore the possibility of using the rank functions, r_A and r_B, or the score function, s_A and s_B, and their variances $d(r_A, r_B)$ or $d(s_A, s_B)$ as diversity measurements respectively. The diversity $d(A, B) = d(r_A, r_B)$ was used in the information retrieval domain [28] and $d(A, B) = d(f_A, f_B)$ in virtual screening and drug discovery [43] and protein structure prediction [22].

(3) **Frame sequences:** In this paper, we applied CFA to each target at each frame, F_i, $i \in [1, f]$. Our performance results and comparisons were based on averaging the MSSD's over all the frames. We will, in future work, explore the diversity *d(A,B)* between a pair of scoring systems (cues, features or systems) across all frames of a tracking sequence (see [22]). This will have to be done off-line on

stored video sequences. However, exploring diversity along this dimension might shed some light on the variation between different cues, features or tracking systems in the long run. Let $F = \{F_1, F_2, \ldots, F_f\}$ be the set of frames in a video sequence. Let A and B be two cues, features or systems in the set of scoring systems $C = \{C_1, C_2, \ldots, C_m\}$. The *diversity score function* defined on F, $s_{(A,B)}(F) = \sum_{j \in N} |f_A(j) - f_B(j)|$, where j is in $N = [1, n]$, $n = |D|$ and $D = \{d_1, d_2, \ldots, d_n\}$ is the set of tracks, and f_A and f_B are the rank-score functions of the scoring systems A and B respectively. It would lead to the diversity rank function $r_{(A,B)}(F)$ if we sort $s_{(A,B)}(F)$ into descending order. The diversity rank-score function $f_{(A,B)}(F)$ is:

$$f_{(A,B)}(j) = (s_{(A,B)} \circ r_{(A,B)}^{-1})(j) = s_{(A,B)}(r_{(A,B)}^{-1}(j))$$

where j is in $[1, f]$. The diversity rank-score function was defined and studied in the CFA framework [14], [22]. Even though this measurement has to be calculated off-line, on a stored sequence of frames, it allows the diversity between two features across all frames to be studied. It is frame independent and may be more accurate when used in subset selection among cues, features or scoring systems for combination and fusion.

8. References

[1] Bar-Shalom, Y. and Fortmann, T.; *Tracking and Data Association*. (1988): Academic Press.

[2] Comaniciu, D., Ramesh, V., and Meer, P., Kernel-Based Object Tracking; IEEE Trans. on PAMI V.25, N.5, May (2003) pp.564-577.

[3] Borghys, D., Verlinde, P., Perneel, C., and Acheroy, M.; *Multi-level Data Fusion for the Detection of Targets using Multi-spectral Image Sequences.* Optical Engineer 37(2) (1998).

[4] Checka, N., and Wilson, K.; *Person Tracking Using Audio-Video Sensor Fusions.* Proceedings of the MIT Project Oxygen Workshop, 2002.

[5] Cox, I.J. and Hingorani, S.L.; *An Efficient Implementation and Evaluation of Reid's Multiple Hypothesis Tracking Algorithm for Visual Tracking*. Int. Conf. on Pattern Recognition (1994) pp.437-442.

[6] Dasarathy, B.V. (Editor); *Elucidative Fusion Systems – An Exposition.* Information Fusion 1 (2001) pp.5-15.

[7] Divok, C., Kumar, R., Naor, M. and Sivakumar, D.; *Rank Aggregation Methods for the Web.* Proc. of WWW10, Hong Kong (2001), pp.613-622.

[8] Elgammal, A., Harwood, D., and Davis, L.S.; *Nonparametric Model for Background Subtraction*. Proc. 6th European Conference on Computer Vision, (2000), pp.751-767.

[9] Fine, I., and Jacobs, R.; *Modeling the combination of Motion, Stereo, and Vergence Angle Cues to Visual Dept.* Neural Computation (1999) **11**: pp. 1297-1330.

[10] Gavrila, D., *The Visual Analysis of Human Movement: A Survey*. Computer Vision and Image Understanding (1999) **73**(1): pp. 82-98.

[11] Haritaoglu, I., Harwood, D., and Davis, L.; *W4: Who, When, Where, What: A Real-time System for Detecting and Tracking People*. 3rd Int. Conf. on Face and Gesture Recog. (1998) pp.877-892.

[12] Ho, T.K., Hull, J.J., and Srihari, S.N.; Decision Combination in Multiple Classifier Systems, *IEEE Trans on Pattern Analysis and Machine Intelligence*, 16(1), (1994) pp.66-75.

[13] Hu, W.; Tan, T.; Wang, L. and Maybank, S.; *A Survey on Visual Surveillance of Object Motion and Behaviors* IEEE Trans. on Systems, Man and Cybernetics, Part C, V.34 , N.3 , Aug. (2004) pp.334 – 352.

[14] Hsu, D.F., Chung, Y.S., and Kristal, B.S.; *Combinatorial Fusion Analysis: Methods and Practice of Combining Multiple Scoring Systems.* In: (H.H. Hsu, editor) Advanced Data Mining Technologies in Bioinformatics, Ideal Group Inc, (2006), pp. 32-36.

[15] Hsu, D.F., Lyons, D.M., and Ai, J.; *Selecting and Evaluating Combinatorial Fusion Criteria to Improve Multitarget Tracking.* 9th Int. Conf on Information Fusion. Florence, Italy, July (2006).

[16] Hsu, D.F., and Lyons, D.M.; *A Dynamic Pruning Strategy for Real-Time Tracking.* IEEE 19th Int.

Conf. on Advanced Information Networking and Applications. Taipei, Taiwan, March (2005) pp.117-124.

[17] Hsu, D.F., Lyons, D.M., Usandivaras, C.; and Montero, F.; *RAF: A Dynamic and Efficient Approach to Fusion for Multi-target Tracking in CCTV Surveillance*. IEEE Int. Conf. on Multisensor Fusion and Integration. Tokyo, Japan; (2003) pp.222-228.

[18] Hsu, D.F., and Palumbo, A.; *A Study of Data Fusion in Cayley Graphs $G(S_n,P_n)$*, Proc. 7th Int. Symp. On Parallel Architectures, Algorithms and Networks (ISPAN'04), (2004), pp. 557-562.

[19] Hsu, D.F. and Taksa, I.; *Comparing rank and score combination methods for data fusion in information retrieval*, Information Retrieval 8(3), (2005) pp.449-480.

[20] Kittler, J., and Alkoot, F.; *Sum versus Vote Fusion in Multiple Classifier Systems.* IEEE Trans. on PAMI (2003) **25**(1): pp. 110-115.

[21] Kuncheva, L.I., *Diversity in Multiple Classifier Systems.* Information Fusion, 6(1) March (2005) pp. 3-4.

[22] Lin, C.Y., Lin, K.L., Huang, C.D., Chang, H.M., Yang, C.Y., Lin, C.T., Tang, C.Y., and Hsu, D.F.; *Feature Selection and Combination Criteria for Improving Accuracy in Protein Structure Prediction*. IEEE Trans. on Nanobioscience ,V.6, N.2, (2007) pp. 186-196.

[23] Loy, G., Fletcher, L.; Apostoloff, N., and Zelinsky, A.; *An Adaptive Fusion Architecture for Target Tracking*. Proceedings of the 5th Int. Conf. on Face and Gesture Recog. Washington DC (2002) pp.261-266.

[24] Lyons, D., Hsu, D.F., Usandivaras, C., and Montero, F.; *Experimental Results from Using a Rank and Fuse Approach for Multi-Target Tracking in CCTV Surveillance.* IEEE Intr. Conf. on Advanced Video & Signal-Based Surveillance. Miami, FL; (2003) pp.345-351.

[25] Nandhakumar, N., and Aggarwal, J.K.; *Physics-based Integration of Multiple Sensing Modalities for Scene Interpretation.* Proc. of the IEEE. V85, N1, Jan. (1997). pp.147-163.

[26] Melnik, O., Vardi, Y., and Zhang, C-H.; *Mixed Group Ranks: Preference and Confidence in Classifier Combination.* IEEE PAMI, August (2004) V26, N8: pp.973-981.

[27] Moore, J.R., and Blair, W.D.; *Practical Aspects of Multisensor Tracking* in: Multitarget-Multisensor Tracking (Eds. Y. Bar-Shalom & W.D. Blair) Artech House (2000) pp.1-76.

[28] Ng, K.B. and Kantor, P.B.; P*redicting the Effectiveness of Naive Data Fusion on the Basis of System Characteristics.* J. of Amer. Society for Information Sci. V.51 N.12 (2000), pp1177-1189.

[29] Press, W., Flannery, B, Teukolsky, S., and Vetterling, W.; *Numerical Recipes in C.* Cambridge University Press 2002.

[30] Rao, N.S.V., *Finite Sample Performance Guarantees of Fusers for Function Estimators*. Information Fusion, Vol.1 No.1, 2000, pp35-44.

[31] Rao, N.S.V., *A Generic Sensor Fusion Problem: Classification and Function Estimation*. In: Multiple Classifier System. (Ed. F. Roli, J. Kittler, T. Windeatt) Springer-Verlag Lecture Notes on Computer Science Vol. 3077, 2004.

[32] Rasmussen, C., and Hager, G.*; Joint Probabilistic Techniques for Tracking Multi-Part Objects*. Proc. Computer Vision & Pattern Recognition. Santa Barbara, CA; (1998) pp.16-21.

[33] Reid, D., *An Algorithm for Tracking Multiple Targets.* IEEE Trans. Aut. Ctrl. Vol. AC-24, No. 6, December (1979), pp. 843-854.

[34] Reisman, P., Mano, O., Avidan, S., and Shashua, A.; *Crowd Detection in Video Sequences.* Symp. on Intelligent Vehicles. Parma, Italy; (2004) June 14-17. pp 66-71.

[35] Roli, F., and Kittler, J.; *Fusion of Multiple Classifiers.* Editorial, Special issue on Multiple Classifiers. Information Fusion. **3**(4) December (2002).

[36] Schrater, P.R., *Bayesian data fusion and credit assignment in vision and fMRI analysis*. SPIE Int. Symposium on Electronic Imaging Vol #5016. Santa Clara, CA; (2003) pp.24-35.

[37] Sharma, R.K., *Probabilistic Model-Based Multisensor Image Fusion*, Ph.D. Diss. 1999, Oregon Grad. Inst. Science & Tech.: Portland, OR.

[38] Snidaro, L., Foresti, G., Niu, R., and Varshney, P.; *Sensor Fusion for Video Surveillance*. 7th Int. Conf. on Information Fusion. Stockholm Sweden, (2004) pp.739-746.

[39] Triesch, J., and von der Marlsburg, C.; *Democratic Integration: Self-Organized Integration of Adaptive Clues.* Neural Computation 13, (2001) pp.2049-2074.

[40] Tumer, K., and Ghosh, J., *Linear and Order Statistics Combiners for Pattern Classification.* In: A. Sharkey Ed., Combining Neural Nets, Springer-Verlag (1999) pp.127-162.

[41] Varshney, P.K., *Special Issue on Data Fusion.* Proc. IEEE (1997) **85**(1).

[42] Xu, L., Krzyzak, A., and Suen, C.Y.; *Method of Combining Multiple Classifiers and their Application to Handwriting Recognition.* IEEE Trans. SMC (1992) **22**(3): pp. 418-435

[43] Yang, J.M., Chen, Y.F., Shen, T.W., Kristal, B.S., and Hsu, D.F.; Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. of Chemical Information and Modeling* 45 (2005), pp 1134-1146.

Appendix A: Target Hypothesis Pruning

There is less variance observed for the crowded tracking scenario because the scoring system has difficulty distinguishing good and bad hypotheses; the correct choice of target is much less clear cut. We are motivated by this to propose that the track hypothesis score distributions are different in these cases. The track hypothesis score histograms h_a and h_b in Fig. A.1 are proposed as typical for scenarios such as those in Fig 4.1(a) and Fig. 4.1(b) respectively (where the vertical ϕ-axis is frequency and the horizontal p-axis is score of a track hypothesis). By definition of h_a we note that approximately $\int_0^1 h_a(p)dp = \int_0^1 h_b(p)dp$. Let p_c be the value of p when h_a and h_b intersect. In our example graphs, p_c is close to 0.5. The histogram h_a reflects that there are fewer hypotheses with good scores (to the right of p_c) than other hypotheses with clearly worse scores (to the left of p_c). On the other hand, h_b has similar numbers of hypotheses with good and bad scores. Based on this proposition, they then show pruning the pool of tracking hypothesis **D** has a much greater effect on the variation in ranks in a crowded tracking scenario (Fig. A.1 (b)) than in a sparse tracking scenario (Fig. A.1 (a)).

The histograms in Fig A.1 are used to derive the rank-score graph associated with the scoring system for each of the two scenarios. The rank of a track hypothesis is related to its score and the score histogram as follows:

$$r(s) = \sum_{x=s}^{1.0} h(x) \cong \int_s^1 h(x)dx$$

The rank functions r_a and r_b for h_a and h_b respectively in Fig. A.1 can be derived in this fashion and graphed against score to yield the rank-score graphs f_a and f_b in Fig. A.2. From the rank-score graphs in Fig. A.2 it can be shown (see [16] for details) that if a score cutoff p_x is used to prune the track hypothesis pool, then as long as $p_x > p_c$ this will produce a greater variation in ranks in the crowded scenario (f_b in Fig. A.2) than in the sparse scenario (f_a in Fig. A.3). This is apparent from Fig. A.2 since f_a has a steeper slope than f_b in the interval $p_x > p_c$ and $f_b^{-1}(p_c) - f_b^{-1}(p_x) > f_a^{-1}(p_c) - f_a^{-1}(p_x)$. As previously mentioned, the variation of the graph of the rank-score function between two experts has impact on whether a rank combination or score combination produces a better result [19]. Hence, these results illustrate that for a crowded scene, the benefit of score based fusions and of rank-based fusions will vary depending on the hypothesis pool pruning threshold. In crowded tracking scenarios, working with rank and score combinations has distinct advantages: the same score cutoff p_x produces more variations in rank in f_b than that in f_a. Working directly with rank combinations can therefore produce a more accurate result.

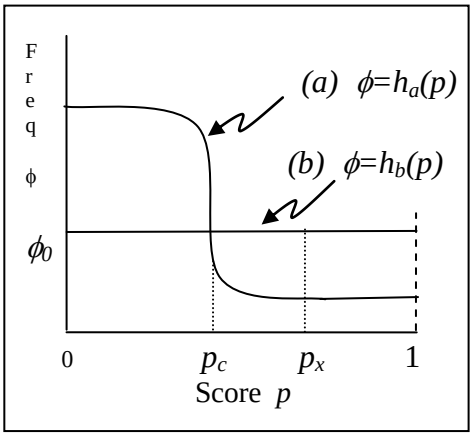


Figure A.1: Frequency of scores for track hypotheses in (a) sparse scenario, and (b) crowded scenario

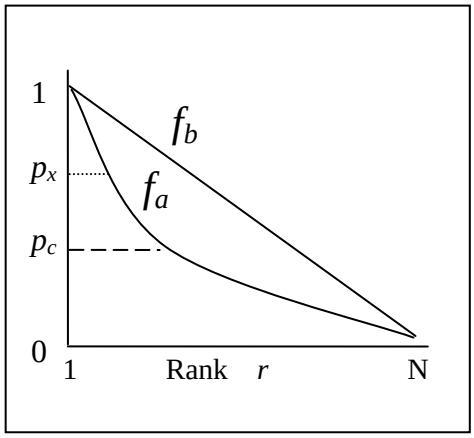


Figure A.2: Rank-score graphs f_a and f_b derived from score histograms h_a and h_b respectively

Appendix B: Feature Selection

For this analysis, we restrict our attention to Fig A.1(b) but now considering *more* than one scoring system. Note that f_b in Fig. A.2 is the typical ideal form of the rank-score graph in this case as related to $h_b(p)$ in Fig. A.1 with the tracking scenario for occluded targets in Fig 4.1(b). However, a given scoring system will vary from this typical case, and may produce a rank-score graph that curves above or below this 'ideal' case. This is shown in Fig. B.1, where h_{b1} and h_{b2} are the histograms for two different scoring systems when tracking in the crowded scenario.

We note again that since $\int_0^1 h_{b1}(p)dp = \int_0^1 h_{b2}(p)dp$ approximately, the up-down curve properties of h_{b1} and h_{b2} have to be opposite. This leads to the rank-score graph of f_{b1} and f_{b2} respectively in Fig. B.2.

The feature selection problem can be phrased as: given the scores for each hypothesis for each feature, which features should be fused to produce the best performing result. We have used the number of false positives (FP) and false negatives (FN) associated with the combination as our criterion for evaluating performance. We show that if scoring systems with *complementary* rank-score functions f_1 and f_2 are combined so that they produce a combination with a rank-score function that is more similar to f_b of Fig. A.2, and Fig. B.2 then this will minimize the false positives (FP) and false negatives (FN) associated with the combination with respect to f_b (see [16]).

A concave-up rank-score graph, such as f_{b1}, assigns fewer ranks to the top scoring tracks and many to the lower scoring tracks, whereas a concave-down rank-score graph, such as f_{b2}, assigns many ranks to the top scoring tracks and few to the lower scoring tracks. We refer to concave-up and down members of this family as complementary graphs. The rank-score graphs for the two scoring systems shown in Fig. B.1 lead to the rank-score graphs shown as f_{b1} and f_{b2} in Fig. B.2, and these are complementary rank-score graphs.

In general, two rank-score graphs won't be perfectly complementary as above, but if the rank-score graph of the combination is closer to the rank-score graph f_b of Fig.A.2, then the FPs or FNs will be

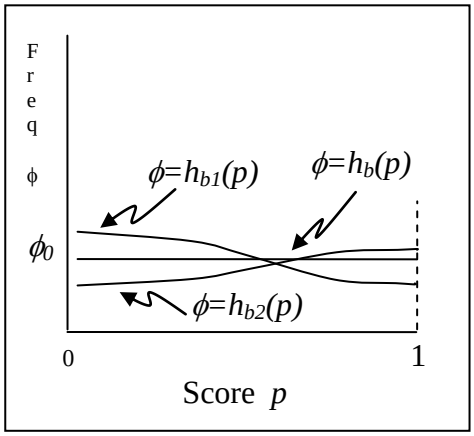


Figure B.1: Histograms for two complementary scoring systems (h_{b1} and h_{b2}) in the crowded scenario.

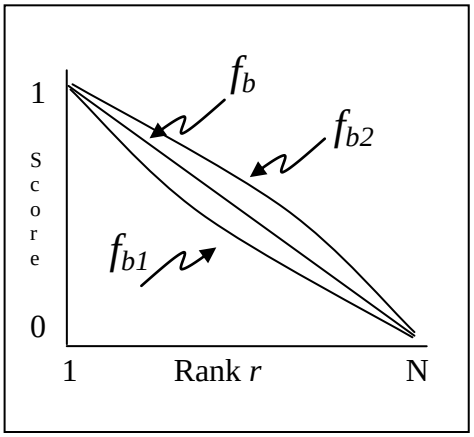
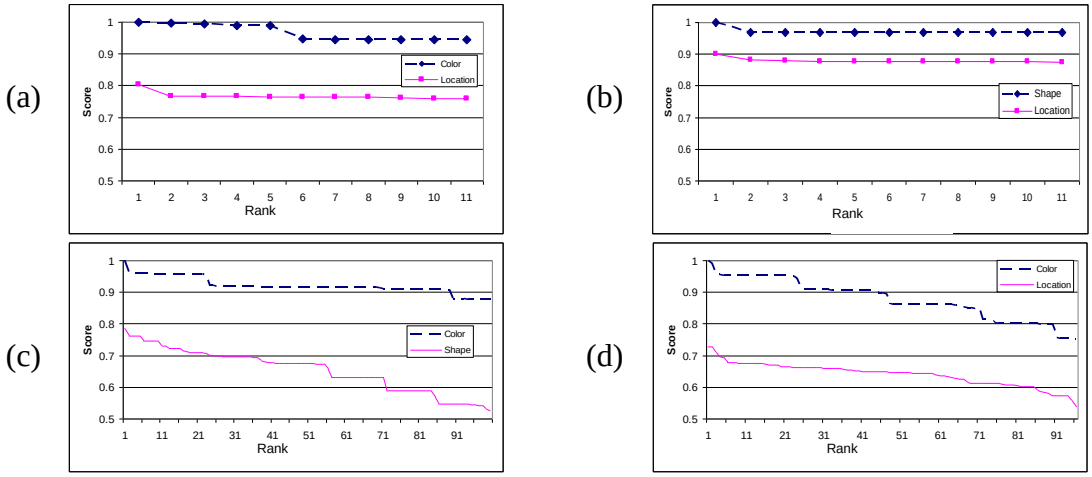


Figure B.2: Rank-score graphs derived from histograms h_{b1} and h_{b2} respectively

reduced. Hence in choosing a subset of features to fuse when tracking in crowded tracking scenarios, selecting features with complementary rank-score graphs will produce a result that minimizes false positives and false negatives.

Appendix C: Example Rank-Score Graphs generated during RUN4.



(a), (c) $\delta_{rs} > 0.05$ (b), (d) $\delta_{rs} \leq 0.05$

Figure C.1: Typical rank-score graphs generated during RUN4 for Sequence 8. The score value is the inverse of the feature similarity value in Section 5.1.

Figure C.1 shows four typical rank-score graphs generated during RUN4. The two of three feature rank-score graphs selected for δ_{rs} (i.e., the ones with largest absolute difference) are shown.