Combinatorial Fusion Criteria for Real-Time Tracking

D. Frank Hsu, Damian M. Lyons and Jizhou Ai
Computer Vision & Robotics Laboratory
Department of Computer & Information Science
Fordham University,
Bronx, NY 10458, USA
{hsu,lyons,ai}@cis.fordham.edu

Abstract

We address the problem of automated video tracking of targets when targets undergo multiple mutual occlusions. Our approach is based on the idea that as targets are occluded, selection of feature subsets and combinations of those features are effective in identifying the target and improving tracking performance. We use Combinatorial Fusion Analysis to develop a metric to select which subset of features will produce the most accurate tracking. In particular we show that the combination of a pair of features A and B will improve the accuracy only if (a) A and B have relative high performance, and (b) A and B are diverse. We present experimental results to illustrate the performance of the proposed metric.

1. Introduction

We propose a novel approach, based on Combinatorial Fusion Analysis (CFA) [7], to automatically select sensory features and fusion operations for recognizing and tracking targets that undergo multiple mutual occlusions. The automated tracking of designated targets in a video image is a general problem that has applications in surveillance, multisensor networks, robotics and virtual reality among other areas. It has, however, remained a difficult problem to solve, especially for tracking targets that undergo occlusion.

Existing work in recognizing and tracking targets that undergo occlusion has focused on modelling the target in such a way that occlusion can be recognized and corrected (e.g., [14]) or modelling the statistical nature of the occlusion (e.g., [29]). As a target moves through a series of occlusions in a crowded, urban space, the method for combination and the process of selecting the set of features/evidence that best identifies and tracks the object changes also. Our principal tool in identifying which features or pieces of evidence are most useful is the emerging field of CFA ([6]-[11], [16], [22], [23], [28]). The use of CFA has three distinct characteristics which are clearly advantages over the existing data and information fusion approaches (see e.g., [4], [26][27]). CFA considers: (A) both score and rank function for each feature/evidence and explores the interaction between the two functions, (B) both combinatorial possibility and computational efficiency of combining multiple scoring systems, and (C) a variety of scoring systems obtained by different methods such as probability, statistics, analysis, combinatorics and computation. In our project, we (a) adopt CFA to inspect and analyse the *full space of possible combinations* of the features or evidence, (b) explore the scoring behavior of each of the features/evidence, (c) propose to use the rank/score function f_A to represent the scoring behavior of a feature or piece of evidence A, and (d) use the difference between the two rank/score functions $d(f_A,f_B)$ to measure the diversity between A and B.

In this paper we present the results of a CFA experiment to select features and feature combinations that improve tracking performance for a video scene of two targets undergoing a series of mutual occlusions. By "improve," we mean that the feature combination yields a target track that is closer to ground truth than either feature yields on its own. The results of this experiment show that a combination of two metrics, a feature performance ratio metric $PR(A,B)$, and a feature rank/score diversity metric $d(f_A,f_B)$, can be used to predict which of the combinatorial fusion alternatives will improve tracking performance at any point.

2. Prior Work

There are several key approaches to target recognition and tracking under occlusions. In the feature-based approach, the crucial problem is determining which features remain robust to occlusion. Lin and Bhanu [15] introduce a feature synthesis strategy for target recognition based on genetic programming approach. Compositions of primitive features are learned that produce improve discrimination between target classes as long as the targets are not overlapping or occluded.

In the contour-based approach, the issue is how to account for the occluded portion of the contour. Koschan et al. [14] use an active shape model approach (ASM) to track the contour of moving human targets. They show that when a target moves from occluded to partially occluded, the ASM contour retains much of its original, unoccluded shape. The ASM is trained using a selection of human

contour shapes, and the ASM is locked onto the target using a manually assigned set of landmark points. This approach handles tracking, but not recognition, and it assumes an initial unoccluded view of the target. Mulayim et al. [21] fuse target color and target texture in a semiparametric statistical model, which they use to define an active contour model in which they can detect a partial occlusion when it occurs, and recover the missing component.

In the model-based approach, since the 'real' appearance of the target (i.e., the model) is known, the issue is to determine how best the occluded view of the target matches its 'real' appearance. Ying and Castanon [29] represent a target as a two dimensional feature template, and they introduce a Markov Random Field approach to modelling an occlusion of the template, and fast algorithms for finding the correspondence between template and target. Liu and Sarkar [17] present a method to fill in the occluded portions of a human target silhouette. They use a Hidden Markov Model to match the target silhouette to a generic stance model, and hence identify the occluded portions.

In a series of papers [8],[12],[18]-[20], the authors have investigated the problem of tracking on a single camera, using multiple feature cue information, in situations where targets engage in multiple mutual occlusions. The core theory for our approach is based on the work of Hsu, Shapiro and Taksa [10]-[11] on characterizing the *scoring behavior*, the relationship between the scores assigned by an expert (e.g., a classifier, a filter, etc.) to a list of candidates and the ranks of the candidates. We have developed a framework, the "Rank and Fuse" (RAF) framework, for fusion for target tracking applications that exploits the combinatorial options for score and rank feature fusion combinations to improve tracking performance. We found that in those situations, a feature fusion operation that included similarity rank information produced a more accurate track than a fusion operation using a Mahalanobis sum of similarities [19]. Our conclusion was that the rank information was less sensitive to the effects of occlusion. However, that research did not provide us with a way to predict *at any frame in the video*, which fusion of which features would provide the more accurate result. A predictor metric needed to be developed that when applied to the video for a specific target, would indicate which features and which fusion operation would yield the most accurate track.

CFA has been applied to a number of areas, including Information Retrieval (IR), pattern recognition (PR), virtual screening (VS) and drug discovery, and protein structure prediction (PSP) ([6]-[7], [9]-[11], [16], [23], [28]). This previous work has suggested that in a combinatorial setting a combination of multiple features improves on the performance over each of those features only if each of the features itself *has good performance* and if the features *are diverse*. In this paper, we quantify performance and diversity, and we present a tracking experiment that explores the performance of rank and score combinations of three features (color, shape and position) to determine the value of the performance ratio and diversity metrics as predictors of the performance of a fusion operation.

3. Combinatorial Fusion Analysis

We consider **each feature measured by a sensor (**which may measure multiple features) **or each piece of the evidence reported by a multiple sensor system** as a scoring system for a tracking and recognition module A on the set of ***n* possible tracks** or **the pool of *n* track hypotheses, *D*** = $\{d_1, d_2,\ldots,d_n\}$. Let $s_A(x)$ be the scoring function which assigns a real number to each d_i in ***D***. We view the function $s_A(x)$ as the score function with respect to the scoring system (feature/evidence) A from ***D*** to ***R*** (the set of real numbers). When treating $s_A(x)$ as an array of real numbers, it would lead to a rank function $r_A(x)$ after sorting the $s_A(x)$ array into descending order and assigning a rank (a positive natural number) to each of the d_i in ***D***. The resulting rank function $r_A(x)$is a function from ***D*** to $\boldsymbol{N}=\{1,2,\ldots,n\}$ (we note that $|\boldsymbol{D}|=n$).

In order to properly compare and correctly combine score functions from multiple scoring systems (multiple features for a single sensor, or multiple items of evidence from multiple sensors) normalization is needed. We simply adopt the following transformation from $s_A(x):\boldsymbol{D}\rightarrow\boldsymbol{R}$ to $s^*_A(x):\boldsymbol{D}\rightarrow[0,1]$ where $s^*_A(x) = \dfrac{s_A(x) - s_{\min}}{s_{\max} - s_{\min}}$, $x \in \boldsymbol{D}$ and $s_{max}= max\{ s_A(x)| x \in \boldsymbol{D}\}$ and $s_{min}= min\{ s_A(x)| x \in \boldsymbol{D}\}$.

Given m scoring systems A_i, $i=1,2,\ldots,m$, with score functions $s_{A_i}(x)$ and rank function $r_{A_i}(x)$, there exist several different ways of combining the output of the scoring systems, including score combination, rank combination, voting, average combination and weighted combination. Initially we will use the average rank (or score) combination as follows. For the m scoring systems A_i with $s_{A_i}(x)$ and $r_{A_i}(x)$, we define the score functions s_R and s_S of the rank combination (RC) and score combination (SC) respectively as:

$$s_R(x) = \sum_{i=1}^{m}\left[r_{A_i}(x) \Big/ m \right], \text{ and}$$

$$s_S(x) = \sum_{i=1}^{m}\left[s_{A_i}(x) \Big/ m \right].$$

As we did before, $s_R(x)$ and $s_S(x)$ are then sorted into ascending and descending order to obtain the rank function of the rank combination $r_R(x)$ and the score combination $r_R(x)$, respectively.

When m scoring systems (features or evidence) A_i, $i=1,2,...,m$, together with the score function $s_{A_i}(x)$ and rank function $r_{A_i}(x)$ 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 $\boldsymbol{D}$ involves sophisticated mathematical, statistical, and computational approaches and techniques (see e.g., [7] and refs). For example, each of the rank functions of the scoring system A_i $i=1,2,...,m$, on $\boldsymbol{D}$, $|\boldsymbol{D}|=n$, can be mapped to a point in the n-dimensional polyhedra called the **rank space**. The n-dimensional polyhedron Q_n is also a Cayley graph with the symmetric group S_n as the vertex set and the adjacency between vertices is defined by a set of generators (a subset of permutations) acting on its vertices.

Remark 1: Previous work in CFA ([6][7], [10]-[11], [13], [16],[23],[28]) in IR, PR, VS and PSP have demonstrated that: (1) the combination of multiple scoring systems (features or evidence) would improve the prediction or classification accuracy rate only if (a) each of the scoring systems has a relatively good performance, and (b) the individual scoring systems are distinctive (or diversified), and (2) rank combinations perform better than score combinations under conditions (a) and (b) and other restrictions.

For the purpose of this paper, our approach considers *combinations of two scoring systems* selected from the $\binom{m}{2}=\frac{m(m-1)}{2}$ possible two combinations using a diversity measure $d(A,B)$ between the scoring systems A and B.

Remark 2: The diversity $d(A,B)$ between A and B has been studied using the score functions $d(s_A,s_B)$ and rank functions $d(r_A,r_B)$ as correlation and rank correlation respectively. The approach of the current proposal is to also use the concept of the rank/score function to measure the diversity between A and B. That is, we include $d(f_A,f_B)$ in addition to $d(s_A,s_B)$ and $d(r_A,r_B)$, where f_A, f_B are the rank/score functions of A and B respectively. The inclusion of $d(f_A,f_B)$ in the measurement of the diversity between scoring systems A and B is one of the novelties of our approach.

When plotting the graph of the rank/score function (hence it is called the rank/score graph) of scoring systems A and B on the same coordinate plane, the diversity measure can be easily visualized. Different diversity measurements have been considered in other application domains ([2], [5]-[7], [10][12], [13], [16], [23], [28]).

Let $s_A(x)$ and $r_A(x)$ be the score function and the rank function of the scoring system A. The rank/score function $f_A(x) : \boldsymbol{N}\rightarrow[0,1]$ is defined as:

$$f_A(i) = (s_A^* \circ r_A^{-1})(i) = s_A^*(r_A^{-1}(i))$$

We note that the set $\boldsymbol{N}$ is different from the set $\boldsymbol{D}$ which is the set of n possible tracks or the pool of n track hypotheses. The set $\boldsymbol{N}$ is used as the index set for the rank function value. The rank/score function so defined signifies the scoring (or ranking) behavior of the scoring system and is independent of the tracks or track hypotheses under consideration. Again, the diversity measure $d(A,B)=d(f_A,f_B)$ can be defined in several different fashions. Here we use the following:

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

4. Experimental Investigation

4.1 Design of Experiment

The goal of the experiment is to determine whether a combination of a feature diversity measure and a relative performance measure are a good candidate metric for predicting whether a fusion of a subset of features will produce more accurate tracking results or not. The RAF tracking software was constructed previously (see [8][12]) to evaluate rank and score fusion operations for multisensory, multitarget video tracking. The three features measured are a color feature, a shape feature and a position feature (see section 4.2 for details). The experiment consisted of running a modified RAF tracking system which calculated and evaluated all feature combinations against ground truth data, and comparing the evaluation results with the values of the proposed predictive metric.

There are 11 possible combinations of the three features and two operations: the basic three features (3), the score combination of any two of these (3), the rank combination of any two of these (3) and the rank and score combination of all three features (2). In this paper, the first 9 of these were evaluated, omitting the combination of all three features since that makes no selection of features.

Evaluation consisted of comparing at each point the top $q=30$ track hypotheses for each target against the ground truth for the video sequence. Ground truth was obtained by having a human observer mark the center of each target in each video frame. Each track is compared to the ground truth by evaluating a Mean

Sum of Squared Differences (MSSD). The performance measure for a combination A, written P(A), is inversely proportional to the average MSSD for the top tracks:

$$P(A) = \frac{q}{1 + \sum_{i \le q} MSSD(track_i)}$$

where $track_i$ is the *ith* ranked track hypothesis for combination A.

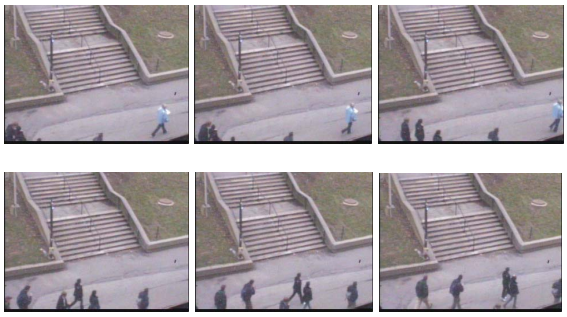

Figure 1: Frames from the test video sequence showing multiple targets and occlusions.

The six 2-combinations are divided into *positive* and *negative* combinations. A combination C that uses features A and B is positive if the performance of C is better than the performance of A and the performance of B, i.e.:

$$P(C) \geq MAX(P(A), P(B))$$

For each combination, two performance metrics are evaluated. The rank-score diversity, calculated for a combination of features A and B as

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

and the performance ratio metric, *PR(A,B)*, calculated as:

$$PR(A,B) = \frac{MIN(P(A), P(B))}{MAX(P(A), P(B))}.$$

On each step, for each combination, the value of $d(f_A, f_B)$, *PR(A,B),*and whether the combination was positive or negative was recorded to a log file.

The video sequence selected for the experiment shows a group of 7 targets that move at varying speeds in a roughly left to right motion (see Figure **1**). The targets undergo repeated mutual occlusions, varying from small occlusions to almost full occlusion.

4.2 RAF Tracker Implementation

In [8][12], [19][20], a multisensor, multitarget tracking system, the RAF tracker, was described. In that tracker, foreground objects are extracted from each frame of the image sequence using a non-parametric background estimation technique. The regions are passed to the three component trackers in the RAF system. Color, location and shape information are collected by applying a tracker-specific measurement to each region c_j in the frame:

(a). Color Tracker: $m_{col}(c_j)$, average normalized RGB color of c_j.
(b). Location Tracker: $m_{loc}(c_j)$, image location of the centroid of c_j.
(c). Shape Tracker: $m_{sha}(c_j)$, area of the image covered by c_j in pixels.

For each frame *i* in the video sequence, a common MHT based [1][3] hypothesis generation module associates these measurements with the set of existing track hypothesis $\mathbf{D}_i$. The gating function is that a track hypothesis be within a standard deviation of the predicted position p_k for target *k:*

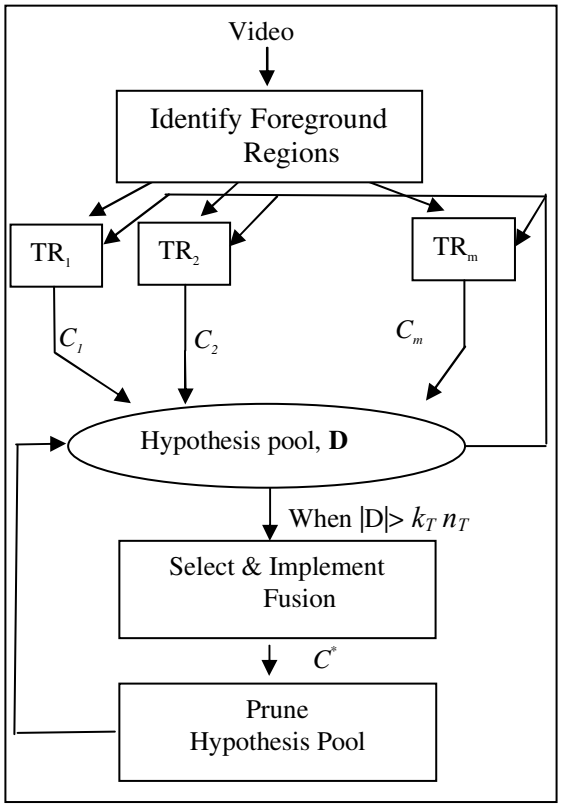

Figure 2: Tracker Block Diagram

$$(p_k - m_{loc}(c_j))^2 < \sigma_k^2$$

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. Each of the three trackers applies its similarity function to determine how well the region fits that target hypothesis. 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*. 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.

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:

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

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 all combinations of two features using both rank fusion and score fusion. The score fusion operation is a Mahalanobis sum (where the coefficients are normalized and inversely proportional to the variance).

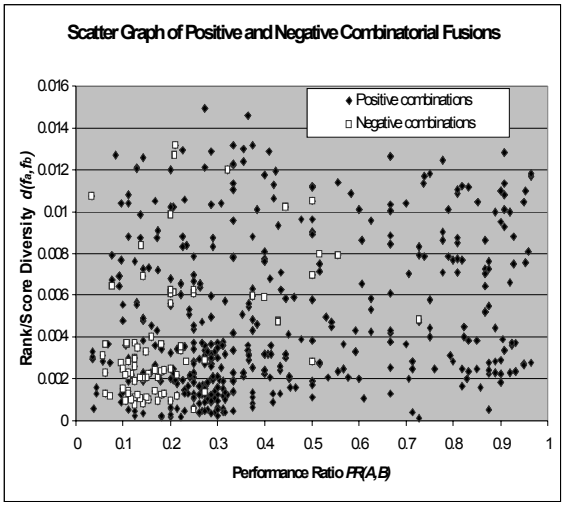

Figure 3: Scatter Graph of Combinatorial Fusion Performance Metrics

4.3 Results

The results are shown normalized in the scatter graph shown in Figure 3. Looking at the graph, it can be seen that the negative combinations, the combinations for which the performance of the combination is worse than the performance of at least one of the combined features, cluster in the lower left, that is, in the area of low relative performance and low diversity. The positive combinations are more evenly scattered through the space, and cluster at a higher relative performance and diversity than the negative combinations.

5. Conclusion

In this paper we have addressed the problem of multitarget tracking in which the targets undergo mutual occlusions. Our approach is based on Combinatorial Fusion Analysis, the investigation of the space of all combinations of measured features. The insight we bring to bear is that as a target moves through occlusions, different combinations of features are necessary to track the target. The specific problem we address is how to select which combinations of which features produce the best tracking performance. We propose and evaluate a combined metric, rank/score diversity and relative performance, for predicting the best combinations.

We conduct a tracking experiment on a video sequence of multiple targets moving together with repeated mutual occlusions. We use the RAF tracking system, developed in previous work, as the basis of this experiment. The system is modified to evaluate all combinatorial options for fusing two of the three features of color, shape and position in a rank and a score combination by comparing the quality of the resultant tracks with a ground truth measurement. Combinations are considered positive if the combination performance is superior to that of either feature, otherwise the combination is considered negative. The values of the rank/score diversity and performance ratio metrics are measured for each case.

Our results show that the negative combinations tend to cluster when graphed in the area of low rank/score diversity and low relative performance. Specifically, we show that the combination of a pair of features A and B will improve the accuracy only if (a) A and B have relative high performance, and (b) A and B are diverse. This indicates that the two metrics proposed can be a useful indicator for to select feature combinations 'on the fly' that improve tracking performance as targets move through multiple occlusions.

References

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

[2] Brown, G., Wyatt, J., Harris, R., and Yao, X.; Diversity Creation Method: A survey and categorization. *Inf. Fusion* 6 (2005), pp5-20.

[3] 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 Recog. (1994) pp.437-442.

[4] Dasarathy, B.V. (Editor); Elucidative Fusion Systems – An Exposition. Information Fusion 1 (200) pp.5-15.

[5] Kuncheva, L., Diversity in Multiple Classifier Systems. *Information Fusion* 6(1), March 2005.

[6] 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.

[7] Hsu, D.F., Chung, Y.S., and Kristel, 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, (2005) in press.

[8] 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 Integratio*n. Tokyo, Japan; (2003) pp.222-228.

[9] 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.

[10] Hsu, D.F., Shapiro, J., and Taksa, I., Methods of Data Fusion in Information Retreival: Rank vs. Score Combination. 2002, *DIMACS TR 2002-58.*

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

[12] Hsu, D.F., and Lyons, D.M., A Dynamic Pruning and Feature Selection Strategy for Real-Time Tracking. *19th IEEE International Conference on Advanced Information Networking and Applications*, March 28-30 (2005) pp. 117-124.

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

[14] Koschan, A., Kang, S., Paik, J., Abidi, B., and Abidi, M., Color active shape models for tracking non-rigid objects. *Pattern Recognition Letters* 24: pp. 1751-1765, July 2003.

[15] Lin, Y., Bhanu, B., Object Detection via Feature Synthesis Using MDL-based Genetic Programming. *IEEE Trans. SMC* 35(3): pp. 538-547. June 2005.

[16] 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 Predictive Accuracy in Protein Structure Classification. *IEEE Symp. On Bioinformatics & Bioengineering* (2005) in press.

[17] Liu, Z., and Sarkar, S., Effect of Silhouette Quality on Hard Problems in Gait Recognition. *IEEE Trans. SMC* Part B. 35(2): pp.170-182, April 2005.

[18] Lyons, D., and Hsu, D.F., Combinatorial Fusion for Target Tracking Using Rank-Score Characteristics. *Sub: Information Fusion 2005.*

[19] Lyons, D., and Hsu, D.F., Rank-based Multisensory Fusion in Multitarget Video Tracking. *IEEE Intr. Conf. on Advanced Video & Signal-Based Surveillance*. Como, Italy. (2005).

[20] 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.

[21] Mulayim, A.Y., Yilmaz, U., and Ataly, A., Silhouette-based 3D Model Reconstruction from Multiple Images. *IEEE Trans. SMC* Part B 33(4): pp. 582-591, August 2003.

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

[23] Ng, K.B. and Kantor, P.B.,; Predicting 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.

[24] Reid, D.B., An Algorithm for Tracking Multiple Targets. *IEEE Trans. on Aut. Control,* 1979. **AC-24**(6): p. 843-854..

[25] 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.

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

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

[28] 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 Inf. & Mod.* 45 (2005), pp 1134-1146.

[29] Ying, Z., Castanon, D., Partially occluded object recognition using statistical models. *Int. J. of Computer Vision* 49(1): pp. 57-78 2002.