Spatio-Temporal Grouping Models and Prominence Models in Perceptual Organization for Semantic Interpretation of Video Shot

Gaurav Harit
gharit@ee.iitd.ac.in
Santanu Chaudhury
santanuc@ee.iitd.ac.in
Electrical Engineering Department, IIT Delhi, New Delhi-110016, India.

Abstract

We focus on the problem of video shot interpretation by making use of perceptual grouping principles on the visual primitives (2D blobs) in a video shot. We present a novel scheme for modeling the homogeneous regions in the form of 2D blobs, that can be tracked easily across the frames. We describe a novel spatio-temporal perceptual grouping scheme, applied on blobs, that makes use of specified temporal consistency model. The grouping results in blob cliques or perceptual clusters or subjects in the scene. A high level semantic interpretation of scenes is done using the principle of Perceptual Prominence of temporal behaviors of the perceptual clusters.

1. Introduction

We develop here an empirical computational model for classifying a video scene into two broad categories: *subject-centric* scenes, that have one or few prominent subjects, and, *frame-centric* scenes, where none of the subjects can be attributed a high prominence and hence the entire frame is a subject of interest. We identify the subjects using perceptual grouping principles in spatio-temporal domain, where we use a temporal consistency model for grouping. The application of temporal consistency or coherence in grouping is motivated from the Gestalt's law of common fate which states that entities having a coherent behaviour in time are likely to be parts of the same organization or group. The associations between visual patterns (primitive entities) in 2-D visual domain are formulated on gestalt principles like similarity, proximity, adjacency, etc. These associations lead to a strong grouping only when, also, consistency is observed in time. There has been prior work [4],[5] on perceptual grouping in spatio-temporal domain using motion cues, but, none of these works attempt to explicitly model temporal coherence. Once the perceptual clusters or subjects in the scene have been identified, we compute the *Perceptual Prominence* [2] of the subjects given a specified set of perceptual attributes, and, a prominence function modelled as a belief network. Perceptual Prominence of subjects is computed after the analysis of the entire scene content and subject behaviours. Unlike visual saliency (in context of bottom up visual attention [3]) which is the saliency or prominence in relation to the immediate sensory experience, Perceptual Prominence deals with the prominence arising as a result of "perceiving", ie, the cognitive interest coming up as a result of the awareness and understanding of the complete visualization space (2D + time).

The next section describes our novel strategy for modeling the homogeneous regions in the form of 2D blobs. Section 3 describes the parameters for our belief network based temporal consistency model. We then describe several prominence models and scene interpretation models and demonstrate their efficacy through experiments.

2. Video Data Clustering

We use K-means clustering on the CIE-LUV color data of the pixels in a video stack consisting of 15 frames for identifying primitive blobs. Unlike [1], we do not include the (x,y,t) coordinates of pixel data to avoid smoothing of segmentation across color boundaries and to reduce computational costs due to high dimensionality. Since using the BIC(Bayesian Information criteria) for deciding the number of clusters in K-means doesn't normally lead to satisfactory clusters (as also noted by [1]), we use a conservative estimate of K (typically 10 to 15) so that the regions of interest get separated out from the background. Once the color planes (given by the cluster centres) for a video stack have been identified, we perform for each frame, an EM (Expectation Maximization) training on the pixels belonging to a particular color plane. The number of Gaussians to be trained depends on the number of significantly large connected components present in the frame for that color plane. Hence, in effect, we model each connected component with a set of 2-D Gaussians. The number of 2D Gaussians required to model a given connected component region is the one that gives the minimum BIC score for EM. Establishing

correspondence between regions across frames requires establishing correspondence between the blobs across frames, within a stack and also across the stacks. The across-frames matching of blobs belonging to a color plane is done based on nearness of position and size parameters of blobs in a particular frame to the subsequent frames in the stack. The best match is taken to be the next instance of a blob. A chain of such instances forms the trajectory of that blob within that stack. Connection of blob trajectories across stacks is done by matching of size and motion properties of a blob in one stack to a subsequent one. When a blob gets occluded for a few frames, it gets tracked when it re-appears again on the basis of its similarity.

3. Spatio Temporal Grouping and the Temporal Consistency Model

The blobs identified in the previous section are the primitive visual patterns considered in our work. These visual patterns can be identified as the nodes N_i of a graph where the links between the nodes represent the association attributes. Since the patterns and the links have a temporal behaviour, the graph is a spatio-temporal graph. The perceptual clusters [2] are identified as cliques in the graph. The grouping probability between a node and a clique is computed using a specified model of temporal consistency. We reproduce here the definition of temporal consistency [2].

Let $S = \{P_1, P_2, P_3, .., P_n\}$ be a set of elementary patterns. Let $\underline{P_i}$ represent a vector of attributes that represent the pattern P_i. As noted above, the node attributes $\underline{P_i}$ and the association attributes $\underline{E}(P_i, C_j)$ (denoting the association between node P_i and clique C_j), must obey the model of temporal consistency. Let $\underline{\mathcal{T}\underline{E}}(P_i, C_j)$ be a vector of attributes that define the temporal behaviour of $\underline{E}(P_i, C_j)$. Let $\underline{\mathcal{T}\underline{P_i}}$ be a set of attributes that define the temporal behaviour of attributes of pattern $\underline{P_i}$. Let $\mathcal{G}(C_j)$ be the grouping measure for the clique C_j that takes into account the *spatio-temporal* attributes of the grouping.

Definition 1 *A temporal consistency model defines a set of temporal behaviour attributes* $\underline{\mathcal{T}\underline{E}}(P_i, C_j)$, $\underline{\mathcal{T}\underline{P_i}}$ *for* $E(P_i, C_j)$ *and* P_i *respectively, and also a function* $\mathcal{Z}$: $\{\underline{\mathcal{T}\underline{E}}(P_i, C_j), \underline{\mathcal{T}\underline{P_i}}, \mathcal{G}(\{P_i\} \cup C_j)\} \rightarrow [0, 1]$ *that gives the grouping probability of a node* P_i *with clique* C_j.

The identification of cliques is done as follows: **Step1**: Form a new Clique if there are nodes that have no clique label. **Step2**: Compute the grouping probability of every node with all currently existing cliques. **Step3**: Relabel the nodes for which the membership probability to an existing clique becomes less compared to another clique. **Step4**: If any node changes its label or there are unlabelled nodes, then go to step 1. The process terminates when the clique labels have stabilized.

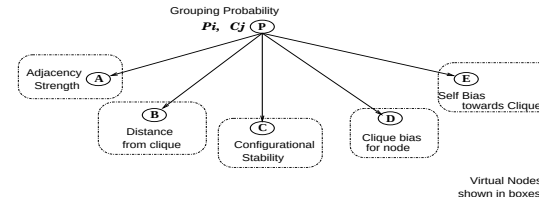


Figure 1. Belief Network for Temporal Consistency.

For our purpose of identifying perceptual clusters in a video scene, we model the temporal consistency function using a belief network shown in Fig 1. In our formulation, we consider that every clique C_j has a *generator*, $P_{gen(C_j)}$, which is the pattern that originates the formation of that clique. $P_{gen(C_j)}$ has the longest lifespan in the clique C_j. We approximate the association measure $E(P_i, C_j)$ to a form $E(P_i, P_{gen(C_j)})$. Hence the association attribute vector $\underline{E}(P_i, C_j)$ now represents the association $\underline{E}(P_i, P_{gen(C_j)})$ between pattern P_i and the clique generator $P_{gen(C_j)}$. We now discuss the parameters (the virtual nodes in the Belief Network shown in Fig 1) of Temporal Consistency model.

3.1. Adjacency Strength

We consider a node to be adjacent to a clique if it has connectivity to the generator of the clique. Computing the adjacency strength of a member to a clique is difficult since a clique member can be connected to the generator through various other clique members depending upon their geometrical arrangement.

The adjacency strength of a node with a clique can be computed as follows: The strength of a link between two nodes N_i and N_j, with number of shared pixels S, can be expressed relative to node N_i or node N_j and is denoted for the two cases as L_{ij} and L_{ji} respectively. Let B_i and B_j be the number of boundary pixels for nodes N_i and N_j respectively. We have $L_{ij} = S/B_i$ and $L_{ji} = S/B_j$. The difference in L_{ij} and L_{ji} depends on the boundary sizes of the two nodes. The initial adjacency strength S_i of each clique member N_i is computed as:

$$S_i = \sum_{\forall j, L_{ij} \neq 0} \min\{L_{ij}, L_{ji}\}$$

However certain links may be at a critical position in the clique, ie, their removal may cause other nodes to get disconnected from the clique. As shown in Fig 2(a), removal of link between N_1 and N_2 may cause the nodes N_2, N_3 to

lose their adjacency to the clique. If this link were having a weak strength then it should limit the adjacency strength of the nodes N_2 and N_3 with the generator. Every link in the clique is removed turn by turn and the adjacency strength S_i of node N_i with the clique is updated as:

$$S_i = \begin{cases} \min\{S_i, L_{km}, L_{mk}\} & \textit{if removal of link between } N_m \& N_k \textit{causes removal of } N_i \\ S_i & \textit{otherwise} \end{cases}$$

When the adjacency strength is more than a threshold (typically 0.2), the node is supposedly considered to have a qualified adjacency with the clique. The overlap strength of a node with a clique is given by the ratio of time instants for which the adjacency strength is more than the threshold, to the total lifespan of the clique.

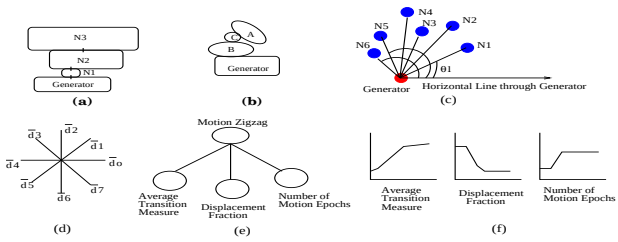


Figure 2. Illustrations for Section 3.

3.2. Clique bias for a Node

The clique bias signifies how much bias or affinity the clique has towards a node. A node may be important for the clique in the sense that it could facilitate the adjacency of other nodes towards the clique, eg, referring to Fig 2(b) , if node C were not a member of the clique, then A and B would get disconnected for the temporal span when C is missing. Hence the clique would apply a bias to C. The clique bias to a node N has been formulated as the sum γ of missed temporal spans of nodes that get disconnected from the clique when the node N is removed. The longer the adjacency period of node with a clique and the more strategic position it has (in terms of affecting the adjacencies of other nodes through its absence), the more bias the clique has towards that node. If α be the sum of temporal spans of all members of the clique, then the clique bias is given by γ/α.

3.3. Self bias of a Node

The Self bias of a node towards a clique is defined as the fraction of the node's adjacency period with the clique lifespan. It signifies the node's affinity towards a clique. Hence even if a node has a small temporal span compared to a given clique's lifespan, but it happens to share an adjacency to the clique for a temporal period that is a large fraction of its life span, then that node will have a self bias towards the clique. We incorporate self bias to account for the nodes/regions that remain mostly occluded in the Perceptual Cluster but show up for a small time span.

3.4. Configuration Stability

There are situations when the temporal consistency model specifies that the relative geometrical configuration of the patterns in the clique should remain stable. To compute the configuration stability, we consider the generator to be the reference node in the clique and compute the orientations and positions of all other nodes relative to the generator. If the relative orientation of a node changes, then it must lead to a low grouping probability of that node with the clique.

Referring to Fig 2(c), the orientation of a node is computed as the angle between the line joining the generator to the node and a horizontal line through the generator. Let ${\theta^t}_i$ be the orientation of node N_i at time instant t, with ${\theta^0}_i$ representing the initial orientation. Let $\delta\theta_i^t = \theta_i^0 - \theta_i^t$, be the change in orientation at time t compared to the initial orientation. If for node N_i, $|\delta\theta_i^t - \delta\theta_j^t| < \xi$ where ξ is the tolerance for angle deviations, typically chosen 10^o, it implies that nodes N_i and N_j have had similar orientation changes. If this holds for N_i and majority of other nodes N_j, then the node N_i is taken to have a stable relative position in the clique. If the node N_i changes its orientation, then $\delta\theta_i^t - \delta\theta_j^t$ will be high for majority of the other nodes N_j. We take the measure of configuration stability for a node as the ratio $1 - (|\delta\theta_i^t - \delta\theta_j^t|)_{avg}/180$.

3.5. Distance from the clique

A node boundary may not share adjacency with a clique, but the two boundaries may maintain a close distance, with the node sharing the same motion as the clique. We compute the distance between a node boundary and a clique boundary as follows. Let p be a point on the boundary of the node. Let d_p be the shortest distance of point p to any of the boundary point of the clique. Arrange the distances d_p for all p in ascending order. The average of the top 15% distances is taken as a measure of the boundary distance of a node to the clique. In order to avoid computing the distance between the boundary of a node to every other emergent clique, we use the distance criteria for the final *clique merging step*, after the clique identification process is over. This allows neighboring cliques to be merged if the distance between them is consistently small for a reasonable period in the temporal overlap of the two cliques. The threshold to be used depends on the sizes of the two cliques and the dif-

ference between the sizes of the two cliques. For two large cliques, the distance threshold is kept very small to avoid them merging. If the distance criteria is satisfied, then the two cliques are merged to form a composite clique.

3.6. Results

The example in Fig 3 shows a few frames of a sampled (5 frames/sec) sequence from a video shot where a person walks down the stairs, turns left to the road, then turns right and walks up the road and finally turns right. The frames are not the original ones, but those obtained after K-means color clustering using 10 centres. The foreground blobs are identified as the ones that show a distinct motion compared to the blobs near the frame boundary. Our assumption that the background blobs would be near the frame boundary holds correct for most of the scenes and allows a reasonable estimation of camera motion. However for scenes where there is no relative motion between the foreground and the background, the foreground blobs have to be interactively marked. Fig 3(A) shows the foreground blobs and the cliques identified. The person initially holds his rucksack at his shoulders and then brings it down as he walks. Fig 3(A) shows the cliques identified using a temporal consistency model where the configuration stability term was not taken into account. Fig 3(B) shows the cliques when the configuration stability term was also used. Since the orientation of the rucksack changes, it gets excluded from the person's clique. Once the cliques are identified, we do its motion analysis.

3.7. Motion Analysis: Epoch Identification

We follow a reasonable assumption that the motion of a perceptual cluster or the clique is described by the motion of the generator node of the clique, ie, the trajectory of the centre of mass of the generator. Given a trajectory, we first identify the epochs where there is a prominent change in the direction of motion. The epochs are the prominent inflection points on the trajectory curve. An epoch is chosen as the point on the curve segment that has the farthest perpendicular distance from the line segment joining the end points of the curve. The subdivision continues recursively for every curve segment till the maximum perpendicular distance of any point on the line becomes less than a threshold.

A motion trajectory may follow an almost uniform direction or a zigzag movement for the lifespan of the clique. We compute the **zigzag measure** of the trajectory in the form of a belief function parametrized on the number of motion epochs, *average transition measure* and *displacement fraction*. We quantize all the directions into 8 prominent directions, $\vec{d_0}$, $\vec{d_1}$,...,$\vec{d_8}$ as shown in Fig 2(d). The gross trajectory length L_i in a given direction $\vec{d_i}$ is the

Figure 3. Illustrative Example

sum of the trajectory lengths from those epoch intervals that have motion in direction $\vec{d_i}$. The transition weight w_{ij} between directions, say, $\vec{d_i}$ and $\vec{d_j}$ is the minimum number of steps required to go from $\vec{d_i}$ to $\vec{d_j}$. Referring to Fig 2(d), w_{05} is 3 and w_{26} is 4. The transition measure of an epoch where the direction changes from $\vec{d_i}$ to $\vec{d_j}$ is given by $w_{ij} \cdot (\min(L_i, L_j)/\max(L_i, L_j))$. The average transition measure for a given trajectory is the average of the transition measures of all the epochs. The displacement fraction is given by the fraction of the displacement, from the start uptill the end point of the trajectory, to the total trajectory length. The belief network and the probability tables showing the dependence of motion zigzag measure on the parameters are shown in Fig 2(e,f).

4. Perceptual Prominence Models for Foreground Subjects

In our previous work [2] we have given a formal definition of Perceptual prominence. Briefly, it is the contextual perceptibility of a cluster under a specified interpretation (or prominence model) that is parametrized on a set of perceptual attributes. For our purpose we have used the following attributes for computing the prominence: motion type,

amount of motion, size, lifespan. For a given scene, the prominence of the subjects can be computed based on the overall shot based characterization of these attributes over the entire life span. This is called the *Global Prominence*. However, if we compute the prominence of a subject for every frame, based on the attributes local to that frame, then it is called the *Local Prominence*. For computing local prominence, we take into account the size of the subject in that frame. Amount of motion (speed) is taken to be the average motion for the epoch interval in which that frame lies. The lifespan is not considered while computing the local prominence. Using different interpretations of Prominence we get different Prominence models as we discuss here.

4.1. Prominence Models

We use a belief network (Fig 4(a)) to model the function $\mathcal{F}$. The nodes are virtual and signify the attributes on which the Prominence model is parametrized. For our purpose, we have formulated 4 different Global Prominence models:
1. No Motion($\mathcal{P}m^1$): Clusters show little or no motion.
2. Motion in a uniform direction, Small lifespan($\mathcal{P}m^2$): The direction of motion mostly remains constant. For such trajectories, the zigzag measure is small. Also, smaller lifespan lead to larger prominence.
3. Motion in a uniform direction, Large lifespan($\mathcal{P}m^3$)
4. Zigzag Motion($\mathcal{P}m^4$): Higher zigzag for the trajectory.

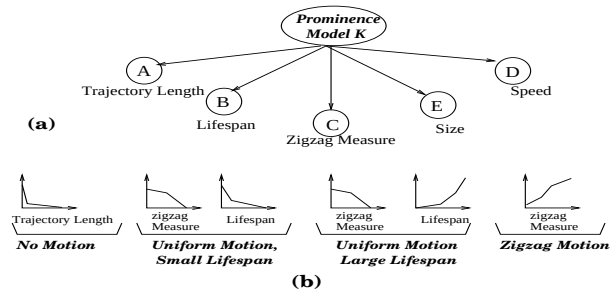


Figure 4. Belief Network for a Prominence Model.

The conditional probability tables used for these prominence models are shown in Fig 4(b).

4.2. Contextual Prominence

The prominence computed for cluster C_j without taking the context into account (ie assuming that only C_j is present), is called its *self prominence* measure $\hat{P}_{c_j}$. Let $\hat{P}^k_{c_j}$ be the self prominence for clusters C_j according to some prominence model k. We omit the superscript k for simplicity. The presence of some other cluster C_i, with a prominence difference $\delta = |P_{c_j} - P_{c_i}|$, affects the prominence of C_j by a penalty factor that is:

$$PF_{ji} = \begin{cases} \delta / \hat{P}_{c_i} & \text{if} \hat{P}_{c_j} < \hat{P}_{c_i} \\ \delta \cdot \hat{P}_{c_i} / (\hat{P}_{c_j} \cdot \hat{P}_{c_j}) & \text{if} \hat{P}_{c_j} >= \hat{P}_{c_i} \end{cases}$$

That is if $\hat{P}_{c_i}$ is larger than $\hat{P}_{c_j}$ then it leads to a larger penalty factor and if it is a smaller, it leads to a smaller penalty factor on $\hat{P}_{c_j}$. The contextual prominence P_{c_j} is given by $\hat{P}_{c_j} \cdot \prod_{\forall i \neq j}(1 - PF_{ji})$.

5. Scene Interpretation

A scene comprises of several perceptual clusters (hence forth referred to as subjects). The *type* of a scene is determined by the composition of the scene (subjects and background) and the dynamics (in behaviour and appearance) of its constituents. We interpret a scene into two broad categories: *Subject Centric* and *Frame Centric*. Subject centric scenes contain fewer subjects with independent spatio-temporal interactions. The subjects normally exist for most of the duration of the shot. Frame centric scenes contain multiple subjects that may be engaged in a group specific activity or independent activities. For our purpose of scene classification, we ignore the implications of the change in the photometric (appearance) properties and consider only the behavioural dynamics of the subjects. Every scene type dictates a specific composition and behavioural model for its constituent subjects. The subjects' perceptual attributes should show adherence to that scene-type-specific behaviour model (henceforth referred to as the scene interpretation model or simply the scene model $\mathcal{S}m$). The behavioural specifications on subjects, dictated by a scene model, can be translated to prominence values of subjects for a set of prominence models. The gross adherence of the subjects to a prominence model $\mathcal{P}m^k$ can be computed as the average P^k_{avg} of the prominence values of all subjects. However, instead of taking into account all the subjects, we consider only a subset of all subjects, $S^k_{high\_P}$, that have the higher values of prominence. We look for the best adherence to model $\mathcal{P}m^k$, rather than the gross adherence. But the question is how many subjects should be considered to compute the *best* average prominence? Instead of relying on any threshold on prominence values to determine the subjects that can be considered in set $S^k_{high\_P}$, we determine the best number of clusters as the one that gives the maximum scene probability accounted *only* by that prominence model (ie given the P^k_{avg} and n^k for the set $S^k_{high\_P}$). The formal algorithm is as follows:

Let $\{ P^k_{c_1}, P^k_{c_2} \ldots P^k_{c_n} \}$ be the prominence values obtained by applying a given prominence model $\mathcal{P}m^k$. We sort the prominence values in descending order, ie, the clusters that gather higher prominence according to the given

model will occur first. Given this prominence vector, we generate a new prominence vector that has, at every position, the average of the prominences upto that position. This cumulative average vector is $\vec{P}^k_{avg} = \{P^k_{c_1}, (P^k_{c_1} + P^k_{c_2})/2), \ldots (P^k_{c_1} + \ldots + P^k_{c_n})/n)\}$.

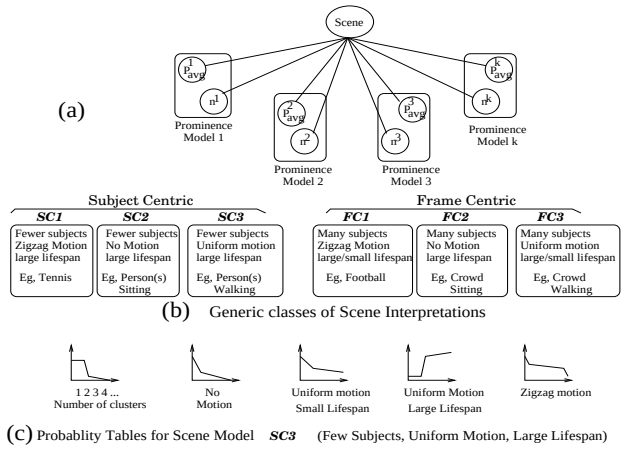


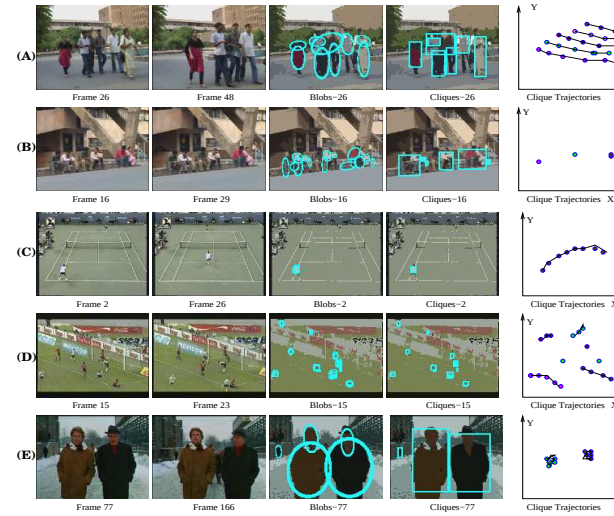
Figure 5. Framework for Scene Interpretation.

The scene model $\mathcal{S}m$ specifies the belief in the scene, given a set of parameters. We parametrize the model $\mathcal{S}m$ on $\{P^k_{avg}, n^k\}$ where P^k_{avg} is one of the average values in $\vec{P}^k_{avg}$ and n^k is the corresponding number of prominence values used for computing P^k_{avg}. We choose only that particular P^k_{avg} that makes the scene probability maximum. The overall belief in a scene is computed using a belief network (Fig 5(a)) where there is a virtual node corresponding to each set $\{P^k_{avg}, n^k\}$ for all the Prominence models applied to the scene.

We further subcategorize our two broad scene categories based on subject motion type: zigzag motion, uniform direction motion and little/no motion. Some generic types of scenes (named *SC1, SC2, SC3, FC1, FC2, FC3*) are listed in Fig 5(b). Conditional probability tables for one of the scene model *SC3* are shown in Fig 5(c).

6. Results and Conclusions

Fig 6 shows five scene examples with the foreground blobs and the cliques identified in those scenes. The trajectories of the subjects(cliques) are also shown. The scene E is of two persons walking. The motion comes due to their wobble as they walk. However, their trajectories are confined to small areas. The table in Fig 6 shows the probabilities with which a given scene affirms to each of the possible scene interpretations. The examples shown are representative of the type of scene they belong to.



Scene Interpretations

	SC1	SC2	SC3	FC1	FC2	FC3
Scene A	0.087	0.008	0.213	0.275	0.037	0.379
Scene B	0.013	0.305	0.077	0.142	0.395	0.069
Scene C	0.516	0.115	0.331	0.019	0.006	0.013
Scene D	0.006	0.111	0.181	0.357	0.041	0.304
Scene E	0.169	0.271	0.391	0.067	0.018	0.084

Figure 6. Results

We have demonstrated the effectiveness of our grouping paradigm and prominence models in obtaining high level scene interpretations. A generic interpretation of a scene is helpful for generating a semantic description of a video shot in terms of its subjects and their interactions or group specific activities. This offers a powerful base for semantic transcoding since the scene interpretation tells us where lies the actual information in the scene.

References

[1] H. Greenspan, J. Goldberger, and A. Mayer. Probablistic Space-Time Video Modeling via Piecewise GMM . *IEEE Trans. PAMI*, 26(3):384 – 396, March 2004.

[2] G. Harit and S. Chaudhury. Video Shot Interpretation using principles of Perceptual Grouping and Perceptual Prominence in Spatio-Temporal Domain . In *proc. ICPR*, volume 4, pages 256 – 259, 2004.

[3] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention.* PhD thesis, California Institute of Technology, Pasadena, California, 2000.

[4] M. Nicolescu and G. Medioni. Perceptual Grouping from Motion Cues Using Tensor Voting in 4-D . In *Proc. ECCV*, volume 3, pages 303 – 308, 2002.

[5] S. Sarkar, D. Majchrzak, and K. Korimilli. Perceptual Organization Based Computational Model for Robust Segmentation of Moving Objects . *Computer Vision and Image Understanding*, 86:141 – 170, 2002.