Integrating and Employing Multiple Levels of Zoom for Activity Recognition

Paul Smith, Mubarak Shah, and Niels da Vitoria Lobo
Computer Vision Laboratory School of Computer Science
University of Central Florida
Orlando, FL 32816
{rsmith,shah,niels}@cs.ucf.edu

Abstract

To facilitate activity recognition, analysis of the scene at multiple levels of detail is necessary. Required prerequisites for our activity recognition are tracking objects across frames and establishing a consistent labeling of objects across cameras. This paper makes several innovative uses of the epipolar constraint in the context of activity recognition. We first demonstrate how we track heads and hands using the epipolar geometry. Next we show how the detected objects are labeled consistently across cameras and zooms by employing epipolar, spatial, trajectory, and appearance properties. Finally we show how our method, utilizing the multiple levels of detail, is able to answer activity recognition problems which are difficult to answer with a single level of detail.

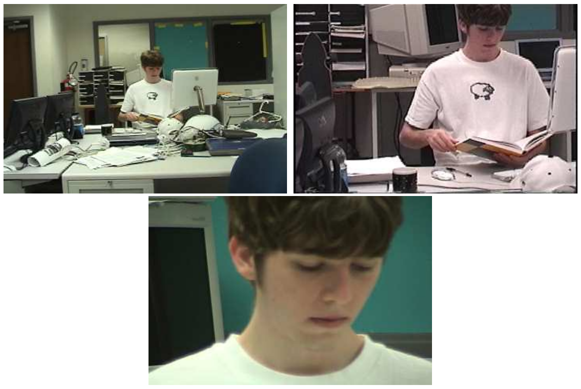


Figure 1: Example of scene showing zoom 1, zoom 2, and zoom 3 views.

1 Introduction

Increasingly many of the activity recognition and surveillance research efforts are utilizing multiple cameras with varying degrees of overlap in the Field of Views (FOV) of the cameras. Single camera systems and even multiple camera systems that are set to the same zoom cannot capture all information available in a scene. We show that a necessary prerequisite for solving many activity recognition problems thoroughly is the introduction of a camera system in which multiple levels of scene detail are employed. In many surveillance environments detailed information on the person's facial features and expressions will need to be integrated with other information such as whether a given person at a computer was the same person who entered the room from a particular door. However, the person first needs to be detected using the low zoom view and tracked from frame to frame. Next his body parts (arms, face, hands, legs) need to be tracked using the medium level zoom view to determine his interaction in the environment. Finally, his facial expression and features need to be analyzed using the fine level zoom. This will provide the basic information needed to perform composite activity recognition. By composite, we mean those activities that are hard to recognize with only one level of detail. In order to achieve the above flow of events, a number of problems need to be overcome. First the camera placement needs to be determined. We experimented with a camera configuration in which there is a hierarchy of $N \geq 3$ zooms which give various degrees of detail in the scene, as shown in Figure 1. There is a large change in zoom and occlusion in the image because we wanted the approach to be robust enough to work without strict camera placement requirements. Another problem is that the head and hands need to be automatically identified in each view and successfully labeled across views. Finally the cameras need to communicate to each other about higher level activities.

Our key contributions are the following: We first demonstrate a bootstrapping process which is able to automatically find the face and hands in the video sequences. Our approach utilizes dynamic color models and multi camera cooperation to achieve better recognition than was possible with independent cameras. Then a method for consistently labeling objects across multiple cameras (each camera having a different zoom) is presented. An improvement over current labeling methods is achieved by incorporating not only epipolar, spatial, and appearance information, but also by developing and integrating a trajectory comparison. Finally it is shown how the individual views can be combined to give better activity recognition capability. We assume that the epipolar geometry of the scene is known, but it could be learned as in [22]. Related work is discussed in Section 2 and some mathematical conventions are given in Section 3. Sections 4 - 6 present the proposed solution. In Section 7 results are discussed and finally we conclude.

2 Related Work

One only has to consider a survey of the activity recognition problem [1] [9] to see the wealth of material available. The problem of integrating multiple levels of detail (MLOD) to improve activity recognition is not as well studied. This paper provides a formulation for studying MLOD in the context of activity recognition.

In [16] multiple cameras are used to cover non overlapping regions to recognize activities, however they do not attempt to use multiple levels of detail to perform finer action recognition. In [7] a method is presented for fusing multiple views, however there is a need for user supervision.

An active vision system is presented in [20] using one static and one Pan-Tilt-Zoom (PTZ) camera to identify and track multiple people. This approach makes a number of restrictive assumptions on the color of people's clothes and number of people present. No activity recognition capabilities are demonstrated.

By combining multiple cameras in an active vision system with stereo vision, [11] is able to perform face and hand tracking and limited gesture recognition. Their correspondence only considers horizontal epipole line information and object size. A multiple camera approach is given in [14] to detect events for an intelligent meeting room, however they do not use the high zoomed cameras for activity recognition. In both these systems the camera positions are known beforehand. We have tried to avoid active vision systems (i.e., PTZ and foveating cameras) in our approach for simplicity.

A key element of any multi camera activity recognition system is the consistent labeling of objects across cameras. An obvious option would be to compute the full 3D alignment using stereo. Basic stereo methods will fail because the assumption of the standard stereo setup is violated [18]. Even after applying polar rectification [17] to our image pairs and then attempting [13] and [3], these direct methods failed. The polar rectification cannot resolve the ambiguities in occlusion and illumination changes across the cameras.

In [19], a feature based method is used, in which the feature point matches are picked randomly. Then a homography is estimated and an error function is minimized which allows the best guesses to help contribute to a better estimate in the next round. In our case however, we do not have a ground plane to work with, which they require, and we have a full 3D scene. As noted in [2] the approach is also sensitive to noise and match ambiguities. Work presented in [5] attempts to find the fundamental matrix and establish trajectory correspondences in 3D scenes. However, their method does not take full advantage of appearance, trajectory, and spatial properties, which we have found adds more robustness to finding the consistent labeling across cameras.

A method is presented in [6] to track across wide field of views. They use epipolar, homography, landmark, apparent height, and apparent color to resolve ambiguities. However the system assumes common illumination across the cameras. We use a better appearance comparison using energy minimization. They neglect to use trajectories themselves, which also provide us a valuable cue to alignment. Further, their approach would have problems without ground plane calibration.

In [15] correspondences are acquired using segmentation and epipolar geometry with information combined from multiple cameras. A ground plane exists to calibrate with in their case. Multiple views with widely different zooms are not considered.

3 Definitions and Conventions

There are many good references on the details of 3D multiview geometry. [10] and [23] provide good introductory knowledge. Only the minimum foundations needed for our purposes are presented here. A pair of images are related by the fundamental matrix, so all points in image I can then be transferred to their corresponding epipolar line in I' by $l = p \cdot F$, where $l = [\ \alpha \quad \beta \quad \gamma\ ]$ are the coefficients of the line equation

$$\alpha \cdot r + \beta \cdot c + \gamma = 0 \tag{1}$$

p is any point in I, F is the fundamental matrix and r, c are the row and column of point p. All epipolar lines will pass through the epipole, which can be found directly from F by taking its singular value decomposition, $F = U \cdot W \cdot V^T$. From here the epipoles are obtained immediately by normalizing the last columns of V and U respectively. To transfer an epipolar line to image coordinates normalize l, then, for lines with slope $|m| > 1$ apply equation 2:

$$p_1 = l \times [\ 0 \quad 1 \quad 0\ ]^T \text{and } p_2 = l \times [\ 0 \quad -1/Y \quad 1\ ]^T \tag{2}$$

where Y is the height of the image and p_1, p_2 are the intersection points of the image with the epipolar line l. A slightly modified operation gives the intersection points for lines with slope $|m| \leq 1$. We will follow the convention that the lowest zoomed image is zoom 1, and the highest zoomed image is zoom 3. The ideas could be easily extended to more than three levels of detail.

Here we present the mathematical notations and conventions we will use in Section 5 in order to more clearly explain the proposed solution. Assume a set of N cameras, with frames $F_{i,j}$ where i is the camera number and j, $1 \leq j \leq J$, is the frame number. Define the set of objects in a particular image frame as $X_{i,j} = \{x_{i,j}^1, \ldots, x_{i,j}^m\}$, with i, j defined as before and m defined as the number of objects in a particular frame. Note, $x_{i,j}^k$ are the four numbers of the bounding box for each object. Thus given a sequence of frames for a particular camera $S_i = \{X_{i,1}, X_{i,2}, \ldots, X_{i,J}\}$ we would like to determine the consistent labeling between all objects in the various sequences $S_i, \forall i$. That is, for a given frame j and object information $T = \{\{X_{1,j}\}, \ldots, \{X_{N,j}\}\}$ expanded as $T = \{\{x_{1,j}^1, \ldots, x_{1,j}^{m_1}\}, \{x_{2,j}^1, \ldots, x_{2,j}^{m_2}\}, \ldots, \{x_{N,j}^1, \ldots, x_{N,j}^{m_N}\}\}$, for each camera we would like to find the mapping

$$(x_{n,j}^k) = \{x_{b_1,j}^{a_1}, x_{b_2,j}^{a_2}, \ldots, x_{b_p,j}^{a_p}\}$$

which takes a particular object k in frame j viewed from camera n, and finds the corresponding object a_k with $1 \leq a_k \leq m_{b_i}$ in camera b_i $\forall i, \ni 1 \leq i \leq N$, $i \neq n$, for frame j, if the object is visible. zooms. We have subscripted m to stress this fact that the number of objects can vary between frames and/or cameras.

4 Detection and Tracking of Heads and Hands

For activity recognition the faces and hands first need to be detected and tracked. We use a boot strapping approach which first finds the head regions and then builds color models of these regions which are used to find the hands. The head regions are detected using the object detector described in [21]. Using the RGB pixel values of the head region, the following function $\forall i, j, k \in \mathbb{Z}, |i|, |j|, |k| \leq N$

$$h(r+i, g+j, b+k) \leftarrow h(r+i, g+j, b+k) + e^{-\left(\frac{i^2+j^2+k^2}{2\sigma^2}\right)}$$

is applied to create a gaussian weighted color model. A similar function $\forall i, j, k \in \mathbb{Z}, |i|, |j|, |k| \leq N$

$$h(r+i, g+j, b+k) \leftarrow h(r+i, g+j, b+k) + e^{-\left(\frac{i^2+j^2+k^2}{2(\sigma-1)^2}\right)}$$

is used to weight the negative samples. In [12] the remaining color pixel values are treated as negative samples. This will not produce a good color model in our case because the hand regions will count as negative samples. To overcome this limitation, after building a color model using the positive sampled regions, this intermediate color model is stored. The final color model is only negatively weighted by those samples which did not show up positively in the intermediate color model. This prevents the hand regions from contributing adversely to the final color model and provides better segmentation. An appropriate threshold can be chosen to make a binary decision $H(r, g, b)$, which can then be used to segment the images.

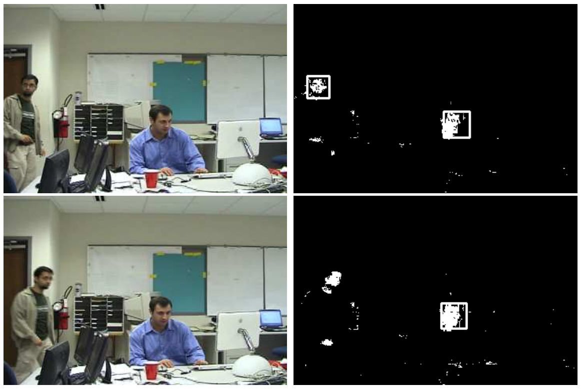


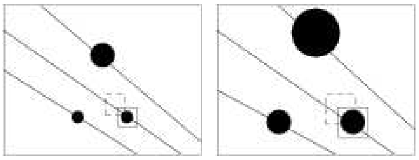
Figure 2: Output from the face detector and color segmentation are shown for zoom 1. Though no explicit color model has been generated for the hands, they show up reliably even for multiple people. Initially the person's head on the left is found, but later in the sequence the face detector misses the correct head(bottom row left side of output image), though the color segmentation still recognizes this region as a skin region.

Since the face detector is for frontal head regions only, the color model will be helpful for detecting hands and heads with small variations in viewpoints. Figure 2 shows the input images on the left and the output of the color segmentation and head detection on the right. Detected heads were drawn with a rectangle around them.

Once a detected head given by the face detector has been present for more than four frames, a mean shift [8] tracker is tracks this head, which will provide further tracking information. Note there is no limitation to how many heads can be in the scene at one time. An alternative approach would be to attach mean shift trackers to head regions whose centroids project to epipolar lines that intersect found head regions in all other views.

Next the hands must be found and tracked in each view. We could simply track all skin colored regions, but this has problems as there are many spurious skin regions marked. Better detection is possible using multiple cameras. First, using the color model, all possible hand candidates are labeled in each sequence. Hand candidates are those that have size$\sum_i H(I(x_i, y_i)) \geq \delta \cdot AverageHeadSize$, where I is an RGB color value and $\delta = .05$. The computation is performed at all levels of detail.

Once all candidate hand regions are labeled, the epipolar



(a) *(b)*

Figure 3: Unambiguous Hand Labeling. In (a) we see three skin regions. The largest one is the head and has already been identified in the first stage. The two smaller dark regions are the hand candidates. The one on the right in (a) is found. Its centroid is projected to the corresponding epipolar line in (b). This line is searched and it is found that there is a hand candidate on this epipolar line.

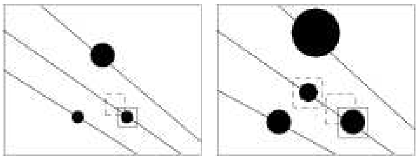
geometry is used to confirm or reject the presence of a hand on an epipolar line in another view. Figure 3(a) is a lower zoomed image, and Figure 3(b) is a higher zoomed image. The lines in each image, are the corresponding epipolar lines from the other image. To do the match across sequences, the epipolar lines are searched for a region with size $\epsilon \cdot AverageHeadSize$. If there are multiple hand candidates along this line, the search is deemed ambiguous, and no hand tracks are introduced. This can be seen in Figure 4. If there is only one valid hand match in the other view, a mean shift tracker is attached to this region in both views, and the hand colored region is tracked across frames. In subsequent frames the color segmentation guides the mean shift tracker. This method is able to successfully detect the face and hands and introduce mean-shift tracks for these regions. When there are multiple head and hand regions and when there are other objects that need to be tracked, the cameras will need to have a consistent set of labels for all objects. A method to establish these consistent labels across cameras is presented next.

5 Establishing Consistent Set of Labels Across Cameras

In order to allow the cameras to communicate object information to one another, a method to determine the consistent set of labels across the cameras needs to be found. For simplicity we will describe our method using two cameras. The ideas can easily be extended to work with additional cameras. Given two cameras, $C_a$and C_b we want to determine the consistent set of labels for objects between cameras for frame j (see Section 3 for a more precise definition).

Our approach has the following components:

- Minimize epipolar line projections for each object



(a) *(b)*

Figure 4: Ambiguous Hand Labeling. In (a) we see three skin regions. The largest one is the head and has already been identified in the first stage. The two smaller dark regions are the hand candidates. The one on the right in (a) is found. Its centroid is projected to the corresponding epipolar line in (b). However we see that there are two hand candidates on this epipolar line. No hand tracks are introduced in this time instant because there is an ambiguity as to which hand candidate in (b) corresponds to the hand candidate in (a).

- Minimize the spatial and trajectory constraints
- Minimize appearance constraints for each object

First, for the j^{th} frame $\forall m$ features: $X_{a,j} = \{x^1_{a,j}, \ldots, x^m_{a,j}\}$ compute the centroids $\mathbf{p_i}$ and make a set of all centroids $\mathbf{P_a} = [x_{c,1}, y_{c,1}], \ldots , [x_{c,n}, y_{c,m}]\}$ in camera C_a. Transfer these points using the fundamental matrix to get the set $\mathbf{A}$ of corresponding epipolar lines $\{l_1, \ldots, l_m\} = \{\{\mathbf{p_1} \cdot \mathbf{F}\}\}, \ldots, \{\mathbf{p_m} \cdot \mathbf{F}\}$ in camera C_b that corresponds to the centroids $\mathbf{P_a}$ from C_a. Apply equation 2 to find the image intersection points of the epipolar lines in the set A.

Then generate a set of centroids $\mathbf{P_b} = \{[x_{c,1}, y_{c,1}], \ldots,$ $[x_{c,n}, y_{c,n}]\}$ in camera C_b $\forall n$ features: $X_{b,j} = \{x^1_{b,j}, \ldots$ $, x^n_{b,j}\}$. There is no requirement for $n = m$. Based on the epipolar constraints, if the i^{th} feature of $C_a, x^i_{a,j}$ is visible in C_b it will lie on the epipolar line l_k. So $\forall p \in \mathbf{P_b}$ and $\forall l \in \mathbf{A}$ the error for this match is the Euclidean distance between the centroid and the epipolar line

$$Err_{p,l} = \frac{|l_\alpha \cdot p_r + l_\beta \cdot p_c + l_\gamma|}{\sqrt{l^2_\alpha + l^2_\beta}} \tag{3}$$

where $Err_{p,l}$ is the error to match the centroid in C_a (whose epipolar line is l) with the centroid in C_b. p_r, p_c are the row and column position of this centroid location. l is the epipolar line with parameters described in equation 1. We can compute the accumulated distance error for every centroid $p \in \mathbf{P_b}$ in C_b with every epipolar line for every frame. To compare objects the accumulated error is averaged over the number of frames that had a valid track for each object. Now since every epipolar line was generated by an object in C_a, we can select for every object, x^j in C_b, the corresponding object x^i, in C_a which generated the lowest distance error $Err_{p,l}$. If the number of objects differ across cameras, then the matching occurs only in the direction with less objects.

It is important for the matching to be commutative, so that x^i in C_a matches x^j in $C_b \Leftrightarrow x^j$ in C_b matches x^i in C_a. The above approach does not meet that criteria when multiple centroids in C_a lie on similar epipolar lines in C_b. The next three constraints provide additional restrictions on matched objects to help reduce the incorrect labelings.

5.1 Spatial Constraints

When multiple centroids in one camera map to similar epipolar lines in another camera, the labels can become incorrect (see Figures 5 and 6). This kind of situation can be detected based on spatial inconsistencies. In Figure 5 the centroid of the hand in zoom 2 lies on its epipolar line in zoom 3, similarly with the right hand. However, in zoom 3, the epipolar line that was generated by the hand in zoom 2 is actually closer to the centroid of the head in zoom 3. Spatially in zoom 2 the hand is below the head. So the spatially lower centroid in zoom 2 matches the spatially higher centroid in zoom 3 though there is a spatial intersection in both views which clearly indicates the correct ordering of the centroids. This condition violates the spatial consistency constraint. Two conditions aid in detecting this inconsistency. In the first case, the bounding boxes of the objects intersect each other, shown in Figure 5. We can assume (since the cameras are arranged in a hierarchical manner), that the spatial ordering of objects is consistent across cameras. If the distances to epipolar lines indicate a spatial inconsistency the match in question is penalized. This is done by adding the Euclidean distance between the two bounding boxes's centroids in C_b to $Err_{p,l}$.

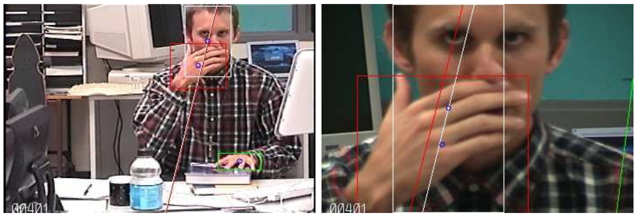


Figure 5: One type of spatial inconsistency. Notice that the head and right hand intersect in both views. In this particular case the epipole is not in the image. However, if it were in the image, the epipolar lines would not be so close. Thus the first spatial constraint tests for intersecting bounding boxes. If the boxes intersect in one view, then intersecting boxes in other views are checked for consistency and penalized if necessary.

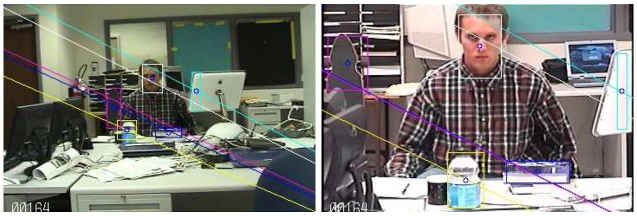


Figure 6: A second type of spatial inconsistency. In this case the bounding boxes of the skateboard and book do not intersect but the epipolar lines are almost on top of one another. This could result in incorrect labeling. The second spatial constraint penalizes label matches that flip flop the order of the centroids.

Figure 6 demonstrates another case in which the best matched objects violate the spatial consistency. The bounding boxes of the skateboard and book do not intersect but the epipolar lines are almost on top of one another, which will result in the epipolar distance minimization selecting the incorrect labels. By considering pairs of epipolar lines which are close in C_b, the objects that they match to in C_b and the original centroids in C_a, which generated the close epipolar lines are analyzed for spatial consistency. If the labels are not spatially consistent the Euclidean distance between the two bounding boxes's centroids in C_b is added to $Err_{p,l}$.

5.2 Trajectory Constraints

Suppose that two objects with similar appearance are on the same epipolar line. If they alternately take turns moving toward each other, this could present problems for the above constraints, but will present no difficulty for a trajectory analysis. Moving objects in one view must match with similarly moving objects in another view. By penalizing candidate trajectory matches which try to match mobile to stationary objects we can effectively eliminate false positives arising from similarly colored objects moving along the same epipolar line. The penalty is computed by multiplying the $Err_{p,l}$ by .00001.

5.3 Appearance Constraints

Previous methods have considered color similarity of objects between views to increase the accuracy of the label assignments. However directly comparing objects in this way can present difficulties especially when the cameras are not color calibrated. Relative color similarity between objects still can give useful information. After applying the previous constraints to all frames, if there are still ambiguous matches (i.e., those objects for which there is not a 1-1 mapping), then collect these ambiguous objects into two lists. The ambiguous objects in C_a are $A = \{x^1_{a,j}, \ldots, x^q_{a,j}\}$ and those in C_b are $B = \{x^1_{b,j}, \ldots, x^q_{b,j}\}$, where q is the number of ambiguous objects. To get the correct matches, find the permutation of superscript indices in B to minimize the relative error:

$$p = \operatorname*{argmin}_{P} \sum_{i=1}^{|A|} [(\frac{1}{M}) \sum_{x \in x^i_{a,j}}^{M} I(x) - (\frac{1}{N}) \sum_{x \in x^{p_i}_{b,j}}^{N} I(x)]$$

Figure 7 shows some results of the labeling. The tracks that are colored the same were matched across views. In Section 4 the method automatically finds the heads and hands. In order to test the accuracy of the labeling algorithm, we have manually introduced bounding boxes around other objects. The algorithm still correctly labels all objects across all views. More results are presented in section 7.

6 Combining Multiple Zooms for Improved Action Recognition

After performing tracking and labeling across cameras, the final step is to use the multiple levels of detail for improved activity recognition. A first natural situation to detect is whether a person has an object in hand or not as the hand is coming to the face. Suppose that we try to determine whether there is an object in the hand and where it came from. If only a view such as zoom 1 is available then this will present several challenges because there is not enough detail in these lower zooms to determine whether the hand had an object in it, and whether it went to the mouth or the ear. In a higher zoomed view such as zoom 3, there is no way to know where the object originally came from in the scene or where and when to look for the object, but zoom 1 and zoom 2 both can share this information with zoom 3. Thus, multiple zooms need to be combined in a manner such that each zoom level answers the questions that it is best able to answer. We show how to combine multiple levels of detail to detect and analyze these composite actions that are difficult to detect with a single level of zoom.

To identify if there is an object in either hand, the hands in zoom 1 and 2 are analyzed for motion by computing F_t. A short term color segmentation is performed on any moving objects, F_t that are not skin in the region of the hand. If significant motion of non-skin colored pixels is found, the epipolar line in zoom 3 corresponding to $l = \mathbf{p} \cdot \mathbf{F}$ where $\mathbf{p}$ is the centroid of the potential object in the lower zoom is found. This gives the possible epipolar line of the object's position in zoom 3. When significant motion is observed in zoom 3 on this line, zoom 2 transfers the color information of the candidate moving object and based on the epipolar geometry the presence of an object is confirmed or denied.

Note that while it is true that the epipolar geometry maps points to lines (for orthogonal, perspective cameras), we can actually do better and predict the exact location of where to look for the object. Since we have an object position in zoom 2, we can find its epipolar line l in zoom 3. Then intersect this line with the image plane, and only look at these intersection points, P, for entering objects, those that appear in this frame for the first time. Note that $|P| \leq 2$ because the images are planar. This reduction in the search space is possible since we know the object is not yet in zoom 3. For instance if no objects appear in zoom 3 at location P, zoom 3 assumes a false positive was observed. This allows for a bad segmentation in zoom 2 to be auto corrected in zoom 3. Now this will not yet eliminate the bad segmentation in zoom 2 but it stops the propagation of the error. Zoom 2 can then be notified of its error.

If the object is confirmed in zoom 3, then segmentation in zoom 3 can proceed since we have a predicted location P, and color model C. By passing location and color information between cameras, we can achieve better object segmentation. This allows early identification of objects in zoom 3. By passing this updated color and spatial information back to zoom 2 we can update its color and spatial parameters for the object in question, which will allow for better segmentation in the lower zooms. Results from our multi camera segmentation have demonstrated that we are able to correctly determine when an object is in the hand and further, when zoom 2 gives an incorrect result the method is able to determine this in zoom 3 and notify zoom 2. Results are shown in Figures 8 and 9. In the first case zoom 2 triggers that an object is present in the hand because the segmentation is not perfectly correct. This can be seen by observing the hole in the segmented skin image. The zoom 3 segmentation is correct and it does not observe any significant motion of non-skin colored objects, thus it overrides zoom 2's decision and notifies zoom 2 of the incorrect segmentation. In the second example, there is a mobile phone being brought to the face. Here zoom 2 identifies an object and alerts zoom 3 to its possible location and color. Zoom 3 then correctly verifies that an object is present.

7 Quantitative Results

The proposed overall method has been formulated in the context of activity recognition for a hierarchy of views. Our method is able to consistently label objects across cameras without the need for ground plane calibration. When the various constraints are combined we achieve 100% accuracy on our test data sets, as presented in Table 1. Data Sets 1-5 were hierarchy of zoom sequences and Data Set 6 was a sequence with partially overlapping FOVs as found in many surveillance papers [4]. The method of determining whether there is an object in the hand using multiple views has been tested on 10 sequences. In 8 sequences the method was successful in determining if an object was present in the

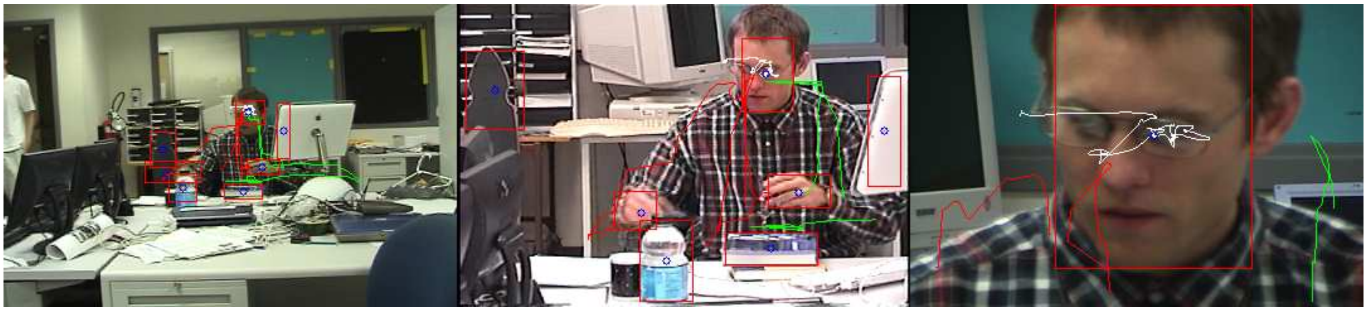


Figure 7: Output of consistent labeling with objects shown in red boxes and the object trajectories superimposed on the last frame in the sequence. The matched trajectories across views are shown in similar colors. All objects were labeled across views correctly.

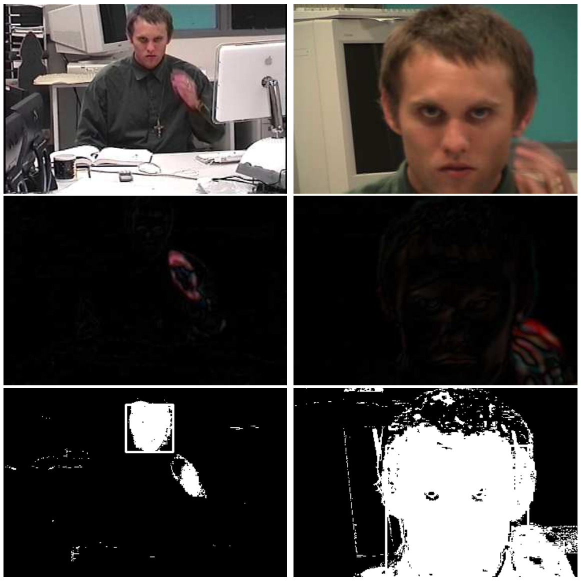


Figure 8: Zoom 2 images are on the right and zoom 3 images are on the left. The first row is the input images. The second row is the F_t images, and the third row is the color segmentation images. In zoom2, a poor color model does not correctly segment all of the hand(bottom left). Thus zoom 2 incorrectly concludes that an object is present. However, in zoom 3, the color segmentation is correct, so zoom 3 can override zoom 2's decision.

hand or not. Some of the cases are very challenging. For instance the method is successful in determining that there is an object in the hands when eye glasses are being brought to the face. With one camera this would be particularly challenging.

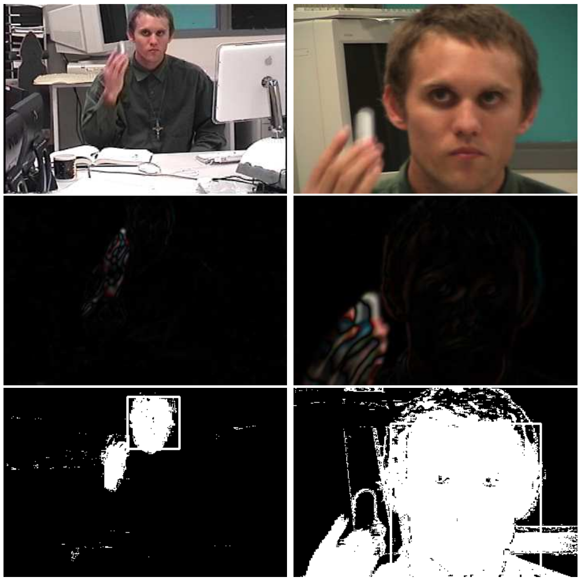


Figure 9: In this case zoom 2 correctly detects an object, and zoom 3 confirms that an object is present. See Figure 8 for more explanation.

Table 1: Consistent Labeling Results

Data Set #	Objects in Camera 1	Objects in Camera 2	Objects in Camera 3	% Matched
1	7	7	3	100
2	7	7	3	100
3	8	8	3	100
4	6	6	2	100
5	7	7	3	100
6	6	6	0	100

8 Conclusion

We have developed a robust method to perform activity recognition. The presented framework is able to combine information over cameras in multiple ways to increase overall system performance. Heads and hands are automatically found and tracked using multiple levels of detail. We have presented a method which is able to incorporate epipolar, spatial, trajectory, and appearance together into a unified framework to achieve consistent object labeling across multiple cameras. The activity recognition module itself is able to integrate multiple levels of detail to determine whether there is an object in the hand in which it would be rather difficult with a single view.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding: CVIU*, 73(3):428–440, 1999.

[2] M. Antone and S. Teller. Scalable, absolute position recovery for omni-directional image networks. In *Computer Vision and Pattern Recognition*, 2001.

[3] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *IEEE Computer Vision and Pattern Recognition Conference*, June 1998.

[4] Q. Cai and J. K. Aggarwal. Tracking human motion in structured environments using a distributed-camera system. In *PAMI*, volume 21, 1999.

[5] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. In *ECCV Vision and Modelling of Dynamic Scenes Workshop*, 2002.

[6] T.-H. Chang and S. Gong. Bayesian modality fusion for tracking multiple people with a multi-camera system. In *Proc. European Workshop on Advanced Video-based Surveillance Systems*, 2001.

[7] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, October 2001.

[8] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-regid objects using mean shift. In *CVPR*, 2000.

[9] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[11] H. Hongo, M. Ohya, M. Yasumoto, Y. Niwa, and K. Yamamoto. Focus of attention for face and hand gesture recognition using multiple cameras. In *Automatic Face and Gesture Recognition*, March 2000.

[12] R. Kjedlsen and J. Kender. Finding skin in color images. In *Face and Gesture Recognition*, pages 312–317, 1996.

[13] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *International Conference on Computer Vision*, July 2001.

[14] I. Mikic, K. Huang, and M. Trivedi. Activity monitoring and summarization for an intelligent meeting room. In *IEEE Workshop on Human Motion*, December 2000.

[15] A. Mittal and L. S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. In *International Journal of Computer Vision*, page 189203, 2003.

[16] N. Nguyen, H. Bui, S. Venkatesh, and G. West. Recognising and monitoring highlevel behaviours in complex spatial environments. In *IEEE Computer Vision and Pattern Recognition Conference*, 2003.

[17] M. Pollefeys, R. Koch, and L. V. Gool. A simple and efficient rectification method for general motion. In *International Conference on Computer Vision*, pages 496–501, 1999.

[18] S. M. Seitz and J. Kim. The space of all stereo images. In *International Journal of Computer Vision*, 2001.

[19] G. P. Stein. Tracking from multiple view points: Self-calibration of space and time. In *DARPA IU Workshop*, pages 1037–1042, 1998.

[20] S. Stillman, R. Tanawongsuwan, and I. Essa. A system for tracking and recognizing multiple people with multiple cameras. GIT-GVU 98-25, Georgia Institute of Technology, August 1998.

[21] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[22] Y. Wexler, A. W. Fitzgibbon, and A. Zisserman. Learning epipolar geometry from image sequences. In *CVPR*, 2003.

[23] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. In *International Journal of Computer Vision*, 1998.