M. Tech. (Computer Science) Dissertation

Human Activity Recognition By Tracking The Global Motion

A dissertation submitted in partial fulfillment
of the requirements for the award of
M.Tech.(Computer Science) degree

By

Apurbaa Mallik

Roll No: MTC1106

under the supervision of

Professor Dipti Prasad Mukherjee

Electronics and Communication Sciences Unit

INDIAN STATISTICAL INSTITUTE

203, Barrackpore Trunk Road

Kolkata - 700 108

Acknowledgements

At the end of this course, it is my pleasure to thank everyone who has helped me along the way.

First of all, I want to express my sincere gratitude to my supervisor Prof. Dipti Prasad Mukherjee for his patience and advice and for the way he helped me think about problems with a broader perspective, I will always be grateful.

I would like to thank all the professors at ISI who have made my educational life exciting and helped me gain a better outlook on computer science.

I would like to thank everybody at ISI for providing a wonderful atmosphere for pursuing my studies. I thank all my classmates who have made the academic and non-academic experience very delightful. It has been great having them around at all times, good or bad.

My most important acknowledgement goes to my family and friends who have filled my life with happiness. Most significantly to my parents who have always encouraged me to pursue my passions and instilled a love of knowledge in me; I am indebted all of them for their endless supply of encouragement, moral support and entertainment.

Abstract

Human Activity Recognition is an active area of research in computer vision with wide scale applications in video surveillance, motion analysis, virtual reality interfaces, robot navigation and recognition, video indexing, browsing,etc. It consists of analyzing the characteristic features of various human actions and classifying them.In a video with static background, activity analysis generally consists of foreground detection, forming the human trajectory ,feature selection and then classifier. However, in real world situation the assumption of static background does not always hold.

Learning global motion patterns from a video is an activity classification problem is important, especially in noisy environment where there is illumination changes,jitters,camera motion and also background is not same for all the videos used for classification.

In our approach we define a method to tackle a real world situation containing illumination changes,jitters and camera motion as an inherited noise in the system. Also a given activity is being performed under different backgrounds. We compute the dense optical flow and quantize it into different labels.Then correct the alignment of the optical flow vectors using probabilistic relaxation labeling in each frame and along the time axis to achieve the dominant motion. We only retain the processed optical flow vectors which are locally maxima.These step removes some amount of noise in the video.It is followed by the construction of the tracks which is a sequence of 3 Dimensional points based on the dominant motion of the system representing the activity.These tracks representing the global motion of the system are not much effected by the induced noise in the videos.We select top dominant tracks of the system based on a criterion which is further processed to represent as the feature descriptor of the given activity.The efficacy of the approach is demonstrated on challenging LIRIS dataset.

Contents

To my Dear Parents.....

Chapter 1

Introduction

Human action recognition is one of the most promising topics in computer vision.The focus of the present work is the automatic recognition of human actions in video sequences.In human action classification, a number of action classes is predefined and, for each class, training samples (positives and negatives) are given. A classifier is then learned from these training samples. Given a test input video sequence, the objective is to issue a corresponding action class label to the entire video sequence. In other words, the question to be answered here is if an action occurs.

By action here we mean a simple motion pattern performed by a single subject ,and in general lasts for a short period of time.

1.1 Motivation

Human action recognition is a very important component of visual surveillance systems for event based analysis of surveillance videos. Visual surveillance systems play a very crucial role in the circumstances where continuous patrolling by human guards is not possible like international border patrolling, nuclear reactors etc. Demand for automatic surveillance systems in civilian applications like monitoring a parking lot, shopping complexes etc. is also increasing heavily. It is difficult and manpower intensive to monitor the data collected from various cameras continuously and this gives rise to the necessity for automatic understanding of human actions and building a higher level knowledge of the events occurring in the scene by the computer vision system.

Recognition of human movements has also been exploited to a large extent for animation like avatar control, for giving gesture based commands to virtual reality interfaces, human computer interactions in smart room like environments etc. Content based video retrieval, indexing and searching is also becoming popular these days . These systems require cognitive vision techniques for analyzing videos which in real life scenarios mostly converges to analyzing human actions in the videos.

1.2 Challenges

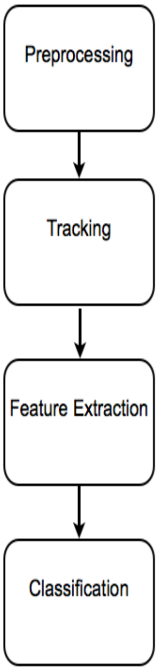

Figure 1.1: *Steps involved in Activity Recognition system*

Figure 1.1 shows the steps involved in Activity Recognition system.

Recognizing Human Action from videos is a challenging problem because physical body motion can look very different depending on the context. For instance similar action with different clothes or in different illumination /background can result in large appearance variation.The same action performed by two different people may look quite dissimilar in many ways.

Implementing real life activity recognition system is a daunting task considering the challenges at each stage of the system like background clutter, dynamic illumination changes, camera movements etc. The action classification problem is characterized by large intra class variability introduced by various sources like the changes in camera viewpoint, anthropometry (body shapes and sizes of different actors), different dressing styles, changes in execution rate of activity, individual styles of actors etc.Due to the following inherited noise in the video/system we do not get a constant background and the general step of background subtraction do not provide good result. The performance of the recognition stage depends on the initial stages and also on the choice of features for action representation.

Figure 1.2 and Figure 1.3 shows the effect of background subtraction when there is no illumination changes.

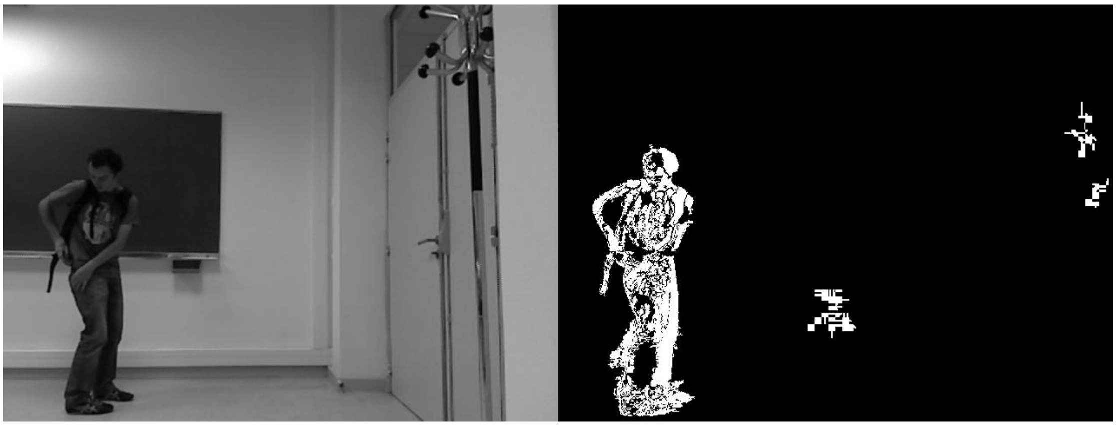

Figure 1.2: Background Subtraction Where There Is No Change in Illumination

Figure 1.4 and Figure 1.5 shows the effect of background subtraction when there is illumination changes.It could not provide good result as there is illumination change in the background due to which we do not get a constant background.

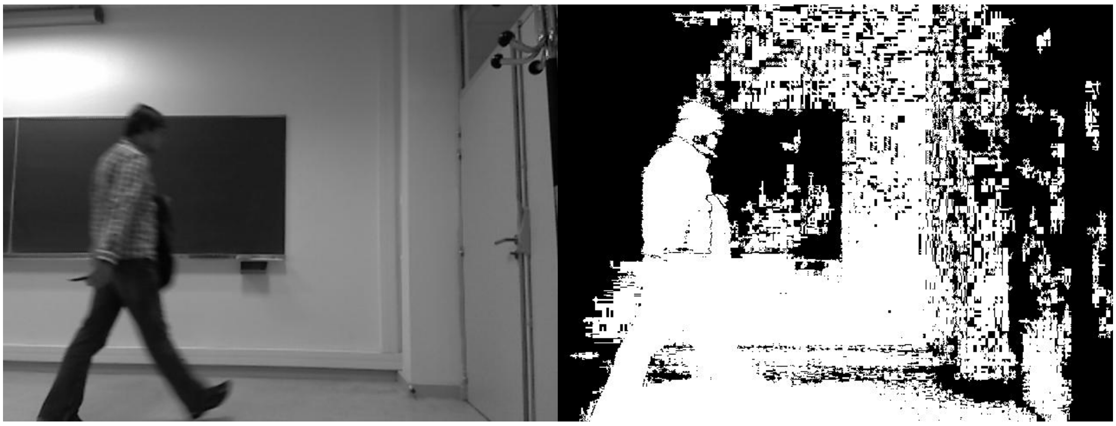

Figure 1.3: Background Subtraction Where There Is Change in Illumination

1.3 Related woks

Human action and activity recognition is an important area of research in the eld of computer vision. A comprehensive review of yhe research has been presented in a number of survey papers [1, 2]. Learning of motion paths or patterns for action recognition has been attempted before in the literarture For instance a method based ons pace-time locally adaptive regression kernels and the matrix cosine similarity measure has been used in [3].[4] used Motion decomposition of lagrangian particle trajectories where as multiple motion trajectories for different body parts has been used in [5].A novel modelling, feature selection and classification architecture for action recogniton can be found in [6] . A non- parametric model for background subtraction is described in [7].Optical flow estimation and their principles [8] has been found very usuful for our approach.[9] has described an approach for finding global motion pattern in complex videos. Combining skeletal pose with local motion for human activity recognition has been used in [10].2 different approaches using kinematic features and bag-of-fatures has been described in [11] and [12] respectively.

1.4 Our approach

We are proposing a method to detect tracks depicting global motion patterns in a non static background. These tracks represents the flow of the respective actions and are processed into direction invariant feature descriptors. Then they are classified into different actions. We have done leave one out cross validation followed by classification using Support Vector Machine.

The main challenge in classification of the dataset containing real world situation is to tackle background clutter, dynamic illumination changes, camera movements etc. Due to large variation between the classes introduced by various sources like the changes in camera viewpoint, shapes and sizes of different actors, different dressing styles, changes in execution rate of activity, individual styles of actors etc.

Figure 1.4 depicts our activity recognition system framework.

For a given video clip of an action firstly we compute the dense optical flow of each frame depicting the pattern of apparent motion of human and objects of interest. We assign a label to every optical flow vector so that they denote a particular octant in an angular radian space. Alignment of optical flow vectors are sometimes hampered by noise.So,for proper alignment of the optical flow vectors it is followed by probabilistic relaxation labeling to achieve dominant motion pattern and the optical flow vectors who has locally maxima magnitude are chosen. These processed optical flow vectors along with their respective labels are then used for construction of tracks which are a sequence of 3 Dimensional points depicting the flow of motion of a point. This is done by minimization of a cost function.We choose top dominant tracks to represent the global motion of the video. These tracks are further processed into direction invariant feature descriptors and classified into various classes.

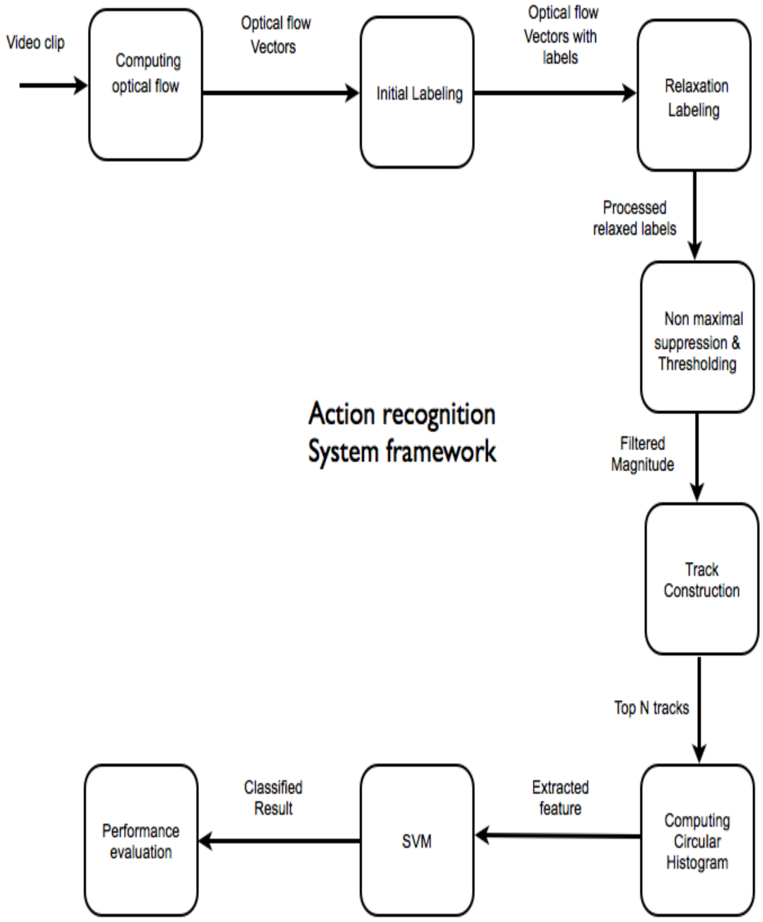

Figure 1.4: *Our Activity Recognition System Framework*

Chapter 2

Optical Flow and Initial Labeling

2.1 Overview

Optical flow or optic flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer (an eye or a camera) and the scene.Sequences of ordered images allow the estimation of motion as either instantaneous image velocities or discrete image displacements.It is in the relation of the motion field. It can be defined as the 2D projection of the physical movement of points relative to the observer to 2D displacement of pixel patches on the image plane.

There are several methods for determining optical flow.Optical flow can be found using Phase correlation which is inverse of normalized cross-power spectrum.Block-based methods uses minimizing sum of squared differences or sum of absolute differences, or maximizing normalized cross-correlation. There exists differential methods of estimating optical flow, based on partial derivatives of the image signal and/or the sought flow field and higher-order partial derivatives.This includes Lucas Kanade method which is regarding image patches and an affine model for the flow field.The Horn Schunck method depends on optimizing a functional based on residuals from the brightness constancy constraint, and a particular regularization term expressing the expected smoothness of the flow field.Buxton Buxton method is based on a model of the motion of edges in image sequences.Another type of differential methods include Black Jepson method dealing with coarse optical flow via correlation.There exists also general variational methods which is a range of modifications/extensions of Horn Schunck, using other data terms and other smoothness terms.Another kind of method includes discrete optimization methods where the search space is quantized, and then image matching is addressed through label assignment at every pixel, such that the corresponding deformation minimizes the distance between the source and the target image. The optimal solution is often recovered through Max-flow min-cut theorem algorithms, linear programming or belief propagation methods.

2.2 Motivation

While an action is occurring it can often described by the flow of motion of the respective object of interest. Optical Flow vectors describe the flow of motion in an action which is very useful in action classification problem. So we have calculated dense optical flow and assign them with labels to represent the magnitude of the motion and the direction in which the motion is taking place at every point of the frame.

2.3 Procedure

2.3.1 Optical Flow Calculation

For a video sequence of some action type A having N frames,for each frame I_f, f $\in$ 1, 2, ...,N is a grey image matrix defined as a function such that for any pixel (i,j), I(i,j) $\in$ ζ, where (i,j) $\in$ Z^2 and $\zeta \in Z^+$ determines the range of the intensity values. For each pair of consecutive frames I_{f-1} and I_f, f $\in$ 2, ...,N we compute the optical flow field $\vec{V}_{f-1}$[paper name].

2.3.2 Assigning Initial Labels

We assign a label L_{ij} to the field vectors of $\vec{V_{ij}}$ of a pixel (i,j) where L = 1,2,3......l. Here each label L_{ij} denotes a particular octant in the angular radian space. We take the value of l as 8, because orientation field is quantized enough when resolved in eight directions in digital grid, i.e., in every 45 degrees.

We can define $\vec{V_{i,j}}$ as optical flow field for any pixel (i,j) where (i,j) $\in Z^2$ as

$$\vec{V_{i,j}} = (V_{x_{i,j}}, V_{y_{i,j}}) \tag{2.1}$$

The labeling process takes place by quantizing $\theta_{i,j}$ as following

$$\theta_{i,j} = \arctan(\frac{V_y}{V_x}) \tag{2.2}$$

After quantization takes place each pixel (i,j) is assigned a label L,where L=1,2,....8 according to the quantized value $\theta_{i,j}$.

If no label corresponds for some value of $\theta_{i,j}$ then it is assigned a no label L_0.

Chapter 3

Relaxation Labeling

3.1 Overview

Relaxation labeling techniques can be applied to many areas of computer vision. The basic elements of the relaxation labeling method are a set of features belonging to an object and a set of labels. In the context of vision, these features are usually points, edges and surfaces. Relaxation labeling has been applied to many problems in computer vision, from edge detection to scene interpretation on the basis of labeled scene components.

In a labeling problem, one is given:

- A set of objects
- A set of labels for each object
- A neighbor relation over the objects
- A constraint relation over labels at pairs (or n-tuples) of neighboring objects

Solution: An assignment of labels to each object in a manner which is consistent with respect to the constraint relation

3.2 Motivation

In real life videos,it contains noise in the background which affects the computed optical flow of the system. So for optical flow to depict a dominant motion we realign them using probabilistic relaxation labeling in each frame and along the flow of time.

3.3 Category

Relaxation Labeling problem can be defined in two categories-

- Discrete Relaxation Labeling
- Probabilistic Relaxation Labeling

3.3.1 Discrete Relaxation Labeling

3.3.1.1 Definition

Early work on scene labeling of employed a discrete relaxation approach in which each scene component was assigned a set of possible interpretations, and inconsistent labeling were removed by examining firstly label pairs on connected segments, and by ensuring secondly that the locally defined consistencies could be linked together in a continuous closed path.

3.3.1.2 Algorithm

3.3.2 Probabilistic Relaxation Labeling

3.3.2.1 Definition

As an alternative to discrete relaxation, probabilistic relaxation allows each scene object to have associated with it not only a set of component labels, but also a weighting assigned to each label in the range (0,1). In general, these weighting are considered

Algorithm 1 Discrete Relaxation Labeling

```
Procedure Labeling
Assign all possible labels to each object
while any object has no label OR no updating possible do
    for each scene object do
        Delete inconsistent labels on the basis of unary constraints
        Delete inconsistent labels on the basis of N-ary constraints
    end for
end while
if any object unlabeled then
    return no solution
else
    return current solution
end if
EndProcedure
```

as probabilities, and so the sum of the label probabilities should be equal to 1. So for each feature, weights or probabilities are assigned to each label in the set giving an estimate of the likelihood that the particular label is the correct one for that feature. Probabilistic approaches are then used to maximize (or minimize) the probabilities by iterative adjustment, taking into account the probabilities associated with neighboring features.

3.3.2.2 Algorithm

Algorithm 2 Probabilistic Relaxation Labeling

```
Procedure Labeling
Define conditional probabilities for each label assignment to each component
while a maximum of an objective function, F is reached OR probabilities cease to
change do
    for each scene object do
        update labeling on basis of compatibility function
    end for
end while
Return current solution
EndProcedure
```

3.4 Procedure

3.4.1 Introduction

Let us assume:

- O is the set { o_1, o_2,.......,o_n } of n object features to be labeled.
- L is the set { l_1, l_2,.......,l_m } of m possible labels for the features

Let $P_i(l_k)$ be the probability that the label l_k is the correct label for object feature o_i .

The usual probability axioms can be applied that:

- Each probability $P_i(l_k)$ satisfies $0 \leq P_i(l_k) \leq 1$ where $P_i(l_k)$=0 implies that label l_k is impossible for feature o_i and $P_i(l_k)$=1 implies that this labeling is certain.
- The set of labels are mutually exclusive and exhaustive. Thus we may write for each i:

$$\sum_L P_i(l_k) = 1 \tag{3.1}$$

Thus each feature is correctly described by exactly one label from the set of labels.

The labeling process starts with an initial, and perhaps arbitrary, assignment of probabilities for each label for each feature. The basic algorithm then transforms these probabilities into to a new set according to some relaxation schedule. This process is repeated until the labeling method converges or stabilizes. This occurs when little or no change occurs between successive sets of probability values.

Generally, a relaxation process will find a local maximum/minimum in the particular criterion, F, defining the quality of match. There are two key issues in relaxation, first the rate of convergence towards the maximum/minimum, and second the position of the local maximum/minimum in its particular optimization space. Convergence can not always be guaranteed, and the algorithm may converge to a local maximum/minimum, based for example on local consistency of several sets of scene objects. To move towards a globally optimum solution, it may be necessary to move through solutions which are locally sub-optimal.

In my procedure I have implemented relaxation labeling using probabilistic relaxation.

3.4.2 Initial Probability Assignment

For a video sequence of some action type A having N frames,for each frame I_f, f $\in$ 1, 2, ...,N is a grey image matrix defined as a function such that for any pixel (i,j), I(i,j) $\in \zeta$, where (i,j) $\in Z^2$ and $\zeta \in Z^+$ determines the range of the intensity values. For each pair of consecutive frames I_{f-1} and I_f, f $\in$ 2, ...,N we have the optical flow field $\vec{V}_{f-1}$. We can define $\vec{V}_{f_{i,j}}$ as optical flow field for any pixel (i,j) where (i,j) $\in Z^2$ in f^{th} frame as-

$$\vec{V_{i,j}} = (V_{x_{i,j}}, V_{y_{i,j}}) \tag{3.2}$$

The magnitude of optical flow field for any pixel (i,j) where (i,j) $\in Z^2$ in f^{th} frame can be defined as the following-

$$m_{i,j} = \sqrt{V_x^2 + V_y^2} \tag{3.3}$$

If $P^1_{i,j}$ denotes the initial probability of a pixel (i,j) of frame f,where f $\in$ 1,2, ...,N-1, then it is calculated by

$$P^1_{i,j} = \frac{m_{i,j}}{\max\limits_{i,j}\ m_{i,j}} \tag{3.4}$$

where $\max\limits_{i,j}\ m_{i,j}$ defines the maximum magnitude value among all the (i,j)pixels of a frame.We now have a set of possible labels $L_{i,j}$, including the no match one, and their initial probability for each frame $P^1_{i,j}$.

3.4.3 Compatibility Coefficients

Intuitively, or from the knowledge of optical flow vectors, one knows that the velocities of two neighboring points constrain each other in that their magnitudes and directions cannot be very different. This constraining relation,is further strengthened the closer the two points are, that is the distance between the two points. The consistency relation, or compatibility, between two neighboring points movement can therefore be determined from the difference between the two velocities and the distance between the two points. In our case, the two points are the pixels of the given frame.

Consider two neighboring pixels p_i, p_j and their respective labels as L_i and L_j. We denote the compatibility coefficient as $C_{i,j}$, to describe the consistency relation between L_i and L_j. In our definition compatibility coefficient, expresses the degree of consistency or inconsistency in terms of our smoothness.In our consideration the compatibility coefficient depends on two parameters-

- The difference between two labels
- The distance between two pixels

First consider the differences between the two labels. We define the compatibility coefficient as

$$\alpha_{i,j} = \cos\ \theta_{i,j} \cdot \left(1 - \frac{\|\ |L_i| - \|\ |L_j|\|}{\max\limits_{i,j}(|L_i|, |L_j|)}\right) \tag{3.5}$$

where, $C_{i,j}$ is the angle between labels L_i, and L_j, $\max\limits_{i,j}(|L_i|, |L_j|)$ is the length of the longer label. The ratio $\left(1 - \frac{\|\ |L_i| - \|\ |L_j|\|}{\max\limits_{i,j}(|L_i|, |L_j|)}\right)$ is the relative difference between the two label's magnitudes, and it takes values within [0,1].

Firstly, when L_i, and L_j are of the same label, the relative difference between their magnitudes is 1, hence

$$\alpha_{i,j} = \cos\ \theta_{i,j} \tag{3.6}$$

In this case, the two labels are most compatible, i.e., $\alpha_{ij} = 1$, if θ_{ij} is 0; or most incompatible, i.e., $\alpha_{ij} =$ -1, if θ_{ij} is Π.

Secondly, when the labels have the same direction, then

$$\alpha_{i,j} = \left(1 - \frac{\|\ |L_i| - \|\ |L_j|\|}{\max\limits_{i,j}(|L_i|, |L_j|)}\right) \tag{3.7}$$

In this case, the two labels are most compatible, i.e., $\alpha_{ij} = 1$, if they are of the same length; or most incompatible, i.e., $\alpha_{ij} = 0$ if the two labels relative difference is 1. In general, the combined effect from label direction and magnitude differences is that the value of α_{ij} is within the range [-1, 1]. $\alpha_{ij} = 0$ indicates that the two labels are independent of each other.

The compatibility relating to the distance d_{ij} between pixels p_i and $p_j$can be expressed as

$$\beta_{i,j} = \exp(\frac{-d_{i,j}}{d_0}) \tag{3.8}$$

where the constant d_0 characterizes flow. In general, do is a function of position. Clearly, like some parameters in many real world problems, do cannot be known beforehand and is also application dependent. For simplicity, it is taken as a constant,and is chosen empirically for our application. Clearly, β_{ij} is in the range [0,1].

The total compatibility coefficient can be simply given by

$$C_{i,j} = \alpha_{i,j} \cdot \beta_{i,j} \tag{3.9}$$

It is noted that for the no-match label we define $C_{ij} = 0$. This is because the no-match case is related to the fact that a vector is unable to be defined, and it is reasonable to assume that this fact is independent of the movement of neighboring points.

3.4.4 Updating Probabilities

If a label,within the label set of a feature, has relatively more support from neighboring features, its chance of being selected as the features' displacement will be enhanced. Its probability will be decreased if the label has relatively less support within the label set.

For each iteration n,we update label L_{ij}'s probability P^n_{ij} ,according to L_{ij}'s consistency relation with the labels of all of neighboring features.This iterative scheme is similar to that given by Rosenfeld and Kak.

The compatibility coefficient C_{ij} can be equivalently seen as the support of label L_j for L_i,weighted by P_{ij} . The total support for L_i therefore, is proportional to

$$q^n_{i,j} = \sum_j C_{i,j} P^n_{i,j} = \sum_j \alpha_{i,j} \beta_{i,j} P^n_{i,j} \tag{3.10}$$

s^n_{ij} is defined as the support function for the n^{th} iteration and is calculated as

$$s^n_{i,j} = \frac{1}{C|\max\limits_r(q^n_{ir})|} q^n_{i.j} \tag{3.11}$$

where $|\max\limits_r(q^n_{ir})|$ corresponds to the largest value of the support found within the label set of p_i, and C≥1 is a constant which controls the speed of convergence as explained in the next paragraph. Clearly, S^n_{ij} is within the range [-1,1], and it can be - 1 or 1 for some values of the j only if C = 1.

When C $\gg$ 1 , all supports $|s^n_{ij}|$ will be small so that their probabilities P^n_{ij}; get modified slowly, i.e.. the system converges slowly. On the other hand, when C is close to 1, some supports S^n_{ij} approach -1. Consequently their associated probabilities P^n_{ij} are suppressed very quickly.

P^{n+1}_{ij} is then updated according to,

$$P^{n+1}_{i,j} = \frac{P^n_{i,j}(1 + s^n_{i,j})}{\sum\limits_r P^n_{i,r}(1 + s^n_{i,r})} \tag{3.12}$$

where P^{n+1}_{ij} is the probability for $n + 1^{th}$ iteration.

3.4.5 Termination Condition

We define the function nz(A) as number of non zero elements in the matrix A. Then we define termination condition on the iterative process using a threshold on the number of non zero elements on the difference between new label and the old label between consecutive iterations, i.e.,

$$\frac{1}{M \times N}\, nz(I^{n+1} - I^{n}) \leq \rho_d \tag{3.13}$$

where ρ_d is the given threshold. When termination condition is met, the maximum probability label within the label set of each feature is taken as an estimate of the features' displacement. If the maximum probability is the no-match labels' probability then the displacement of the feature is undetermined.

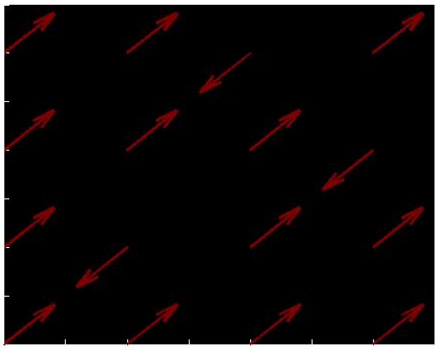

FIGURE 3.1: Before Relaxation Labeling

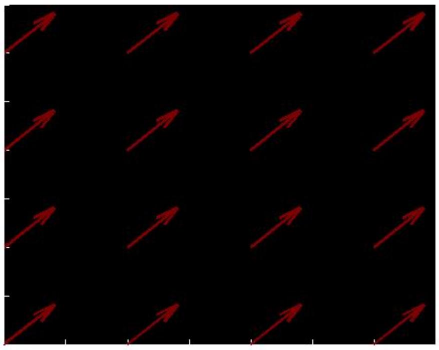

FIGURE 3.2: After Relaxation Labeling

Chapter 4

Non Maximal Suppression

4.1 Overview

Non-maximum suppression is often used along with edge detection algorithms. The image is scanned along the image gradient direction, and if pixels are not part of the local maxima they are set to zero. This has the effect of suppressing all image information that is not part of local maxima.

We will use non maximal suppression in our algorithm to determine if the optical flow magnitude assumes a local maximum in the optical flow direction.

4.2 Motivation

To reduce the effect of noise in the system we have computed the locally maxima optical flow vector along a direction and thresholded it with the mean of magnitude of processed optical flow vectors.

4.3 Procedure

4.3.0.1 Calculation Of Local Maxima

We have already assigned a label L_{ij} to the field vectors of $\vec{V_{ij}}$ of a pixel (i,j) where L = 1,2,3......l. Here each label L_{ij} denotes a particular octant in the angular radian space. As we have taken the value of l as 8, the orientation field is quantized enough when resolved in eight directions in digital grid, i.e., in every 45 degrees.So when L=1 it denotes 0 degree,when L=2 it denotes 45 degree, when L=2 it denotes 90 degree and so on.

m_{ij} is the magnitude of the optical flow vector $\vec{V_{ij}} = (V_{x_{i,j}}, V_{y_{i,j}})$ at pixel p_{ij} calculated as

$$m_{i,j} = \sqrt{V_x^2 + V_y^2} \tag{4.1}$$

So,the process is carried out as

- if the label L_l corresponds to 0 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x+1,y}$ and $p_{x-1,y}$
- if the label L_l corresponds to 45 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x+1,y+1}$ and $p_{x-1,y-1}$
- if the label L_l corresponds to 90 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x,y+1}$ and $p_{x,y-1}$
- if the label L_l corresponds to 135 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x-1,y+1}$ and $p_{x+1,y-1}$
- if the label L_l corresponds to 180 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x-1,y}$ and $p_{x+1,y}$
- if the label L_l corresponds to 225 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x-1,y-1}$ and $p_{x+1,y+1}$

- if the label L_l corresponds to 270 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x,y-1}$ and $p_{x,y+1}$
- if the label L_l corresponds to 315 degrees its magnitude m_{ij} will be considered if it is greater than the magnitudes at pixels $p_{x+1,y-1}$ and $p_{x-1,y+1}$

The new maxima magnitude we get after the above process is denoted by M_{ij} for each frame f=1,2,....N.

4.3.0.2 Thresholding On Mean Of Maxima Magnitude

We define μ as the mean of the maxima magnitude M_{ij}. It can be expressed as

$$\mu = \frac{1}{N}\sum_{ij} M_{ij} \tag{4.2}$$

So the new magnitude $M^{'}_{i,j}$ after thresholding can be expressed as

$$M^{'}_{i,j} = \begin{cases} 0 & \text{if } M_{i,j} < \mu \\ M_{i,j} & \text{otherwise} \end{cases} \tag{4.3}$$

Chapter 5

Tracking The Global Motion

5.1 Overview

Learning dominant motion patterns or activities from a video is an activity classification problem is important, especially in noisy environment where there is illumination changes,jitters,camera motion and also background is not same for all the videos used for classification. Here tracking of human and object are hard if not impossible because of the noise associated with the video. We use instantaneous motion field of the video marked with relaxed labels for learning the motion patterns.

The motion field is a collection of independent flow vectors detected in each frame of the video where each flow is vector is associated with a spatial location. A motion pattern is then defined as a group of flow vectors that are part of the same physical process or motion pattern that is motion of a person/object.

5.2 Motivation

In a low density scene activity analysis generally consists of foreground detection, forming the human trajectory ,feature selection and then classifier. However, in real world situation the assumption of low density does not always hold. In noisy environment background subtraction may be often misleading as the background is not constant due to inherited noise of the system. The motion field is highly effected by the noise like illumination changes, camera motion and jitters which may hamper the detection of the object of interest. So to remove the effect of noise in the system we find the tracks based on the dominant motion of the system representing the activity.

These tracks representing the global motion of the system are not much effected by the induced noise in the videos. The motion flow field is obtained by first using the existing optical flow methods to compute the optical flow vectors in each frame,then properly aligning the optical flow vectors to mark the major motion of the spatial area using relaxation labeling and then combining the relaxed optical flow vectors from all frames of the video into top N tracks representing the global motion field.

5.3 Procedure

5.3.0.3 Estimation of Tracks In The Video

In our context we define track as a sequence of (x_t,y_t) points where t=1,2,.....M and M represents the number of frames in the video representing a motion in the system.

For a video containing M number of frames let each frame be represented by F_t where t=1,2,....M .Each pixel of F_t , $p_{t_{ij}}$ at time=t contains processed optical flow field $V_{t_{ij}}$. The magnitude $M^{'}_{t_{i,j}}$ at $V_{t_{ij}}$ may be zero or may have some positive quantity. Similarly for every pixel in F_{t+1}, we have a processed optic flow vector, say $V_{t+1_{ij}}$ having magnitude $M^{'}_{t+1_{i,j}}$ as zero or some positive quantity .

Considering there is no significant high inter frame motion we can establish between $V_{t_{ij}}$ and $V_{t+1_{hk}}$ where pixel $p_{t+1_{hk}}$ is one of the 8 neighbors of the $p_{t_{ij}}$ when $p_{t_{ij}}$ is placed at the location $p_{t+1_{ij}}$. This correspondence is established by minimization of cost function between the source point and its neighbor based on the following criteria-

- The two points $p_{t_{ij}}$ and $p_{t+1_{hk}}$ should be spatially close to each other.
- The absolute difference between the magnitude $M^{'}_{t_{ij}}$ and $M^{'}_{t+1_{hk}}$ should be minimized.

- The angle of optical flow vector $a_{t_{ij}}$ and $a_{t+1_{hk}}$ at $p_{t_{ij}}$ and $p_{t+1_{hk}}$ should be close to each other.

So we define the cost function C_{hk} where hk is the 8 neighborhood of ij as

$$C_{h,k} = \frac{|M^{'}_{ij} - M^{'}_{hk}|}{\max\limits_{hk} |M^{'}_{ij} - M^{'}_{hk}|} + \alpha * \frac{1 - \cos\,(a_{ij} - a_{hk})}{\max\limits_{hk}\,(1 - \cos\,(a_{ij} - a_{hk}))} \tag{5.1}$$

where $0 \leq \alpha \leq 1$ and $0 \leq \frac{|M^{'}_{ij} - M^{'}_{hk}|}{\max\limits_{hk} |M^{'}_{ij} - M^{'}_{hk}|} \leq 1$.

As value of -1 $\leq$ cos $\theta \leq$ 1,so value of $0 \leq \frac{1 - \cos\,(a_{ij} - a_{hk})}{\max\limits_{hk}\,(1 - \cos\,(a_{ij} - a_{hk}))} \leq 1$.

The minimum cost function among all the neighborhood points p_{hk} with respect to the source point p_{ij} is calculated by

$$C^{'}_{i+1,j+1} = \min_{hk} \frac{C_{h,k}}{\max\limits_{hk}\ C_{h,k}} \tag{5.2}$$

The point $p_{i+1,j+1}$ corresponds to one of neighborhood points p_{hk} which has the minimum cost function with respect to the source point p_{ij} and added as the next coordinate of the track.If there is any conflict ,that is if the cost function between the point and the neighboring point is minimum but equal then we will resolve the the conflict by choosing the neighboring point for which $|M^{'}_{ij} - M^{'}_{hk}|$ value is minimum. We will get multiple number of tracks for a video,each track represented as a series of 3-Dimensional points.

The entire analysis is repeated by taking the size of pixel as $4 * 4$ block, $8 * 8$ block, $16 * 16$ block and $32 * 32$ block. So we have five series of tracks

- for each pixel of the image
- for each $4 * 4$ block of the image
- for each $8 * 8$ block of the image
- for each $16 * 16$ block of the image
- for each $32 * 32$ block of the image

5.3.0.4 Selection of Top Dominant Tracks

After estimation of tracks in the video we get a series of tracks represented by a sequence of 3-Dimensional coordinates. Each of these tracks represents the motion of the initial point in consideration along the length of the video.

We express $Track_i$ represented by series of points (x_1^i,y_1^i), (x_2^i,y_2^i),........(x_M^i,y_M^i) dominant over $Track_j$ represented by series of points (x_1^j,y_1^j), (x_2^j,y_2^j),........(x_M^j,y_M^j) if $dist_i$ $>$ $dist_j$ where $dist_i$ is represented as

$$dist_n = \sum_m \sqrt{(x_{m+1}^n - x_m^n)^2 + (y_{m+1}^n - y_m^n)^2} \tag{5.3}$$

where m=1,2,3....M-1.

Based on the above criterion from the series of tracks,we choose N top dominant tracks. These dominant tracks represents the global motion of the video depicting the action.Here in our experiment we take N=5;

Chapter 6

Computation Of Feature Descriptor

6.1 Overview

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.The main goal in feature selection in supervised learning is to find a feature subset that produces higher classification accuracy.

6.2 Motivation

Feature characterizes its corresponding action and is used for classifying different videos into particular class.

6.3 Procedure

From the last stage we got top N dominant tracks represented by series of 3-Dimensional points (x_1^i,y_1^i),(x_2^i,y_2^i),.......,(x_M^i,y_M^i) where there are M number of frames in the video and i=1,2,....,M.We employ the construction of a direction invariant feature descriptor from these tracks to represent an action.

We describe the vectors of the track as $\vec{T_p} = (\vec{T_{x_p}}, \vec{T_{y_p}})$ where p =1,2,3,......M-1. The vectors are computed as

$$\vec{T_{x_p}} = (x_{p+1} - x_p)\hat{\imath} \tag{6.1}$$

$$\vec{T_{y_p}} = (y_{p+1} - y_p)\hat{\jmath} \tag{6.2}$$

The angle between the vectors represent the change of direction made while an activity is taking place.This is direction invariant feature for an action and do not depend on on the starting and ending location of object of interest in different videos for the same action. It is calculated as

$$\theta_p = \arccos \frac{\vec{T_{x_p}} \cdot \vec{T_{y_p}}}{|\vec{T_{x_p}}||\vec{T_{y_p}}|} \tag{6.3}$$

We distribute the values of θ_p where p=1,2,3.....M-1 in an L-bin histogram. Here each bin denotes a particular octant in the angular radian space. We take the value of L as 16, that means it is quantized in every 22.5 degrees.The histogram H = h_1, h_2, ..., h_L construction takes place by quantizing θ_{xy} and adding up $M_{xy}{}'$ to the bin indicated by the quantized θ_{xy} where p_{xy} is the pixel of the frame in consideration and $M_{xy}{}'$ is its corresponding optical flow magnitude. In mathematical notation,

$$h_i = \sum_p \begin{cases} M_{xy} & \text{if } \theta_p \in i^{th} \; octant \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

So,each video produces one 16-bin histogram leading to 16-dimensional histogram vector which depicts the characteristics of that action. These feature descriptors are used for classification of the action.

Chapter 7

Experiment and Result

7.1 Dataset

The dataset used is the LIRIS human activities dataset which consists of 10 classes. Each of classes can be a normal activity, a human-human interaction or a human-object interaction:

The Activities are

- Activity 01 :: Discussion of two or several people
- Activity 02 :: A person gives an item to a second person
- Activity 03 :: An item is picked up or put down
- Activity 04 :: A person enters or leaves an office
- Activity 05 :: A person tries to enter an office unsuccessfully
- Activity 06 :: A person unlocks an office and then enters it
- Activity 07 :: A person leaves baggage unattended
- Activity 08 :: Handshaking of two people
- Activity 09 :: A person types on a key-board

- Activity 10 :: A person talks on a telephone

This dataset is used to extract training features and test features . We have used 7 video sequences for each video and each video having 200-250 frames at an average. All images have size 480*640. The dataset contains background clutter, dynamic illumination changes, camera movements etc. Large variation between the classes introduced by various sources like the changes in camera viewpoint, shapes and sizes of different actors, different dressing styles, changes in execution rate of activity, individual styles of actors etc increases the classification difficultly level.

7.2 Parameters

We have quantize the dense optical flow into 8 bin histogram so that it is resolved in eight directions in digital grid, i.e., in every 45 degrees. In relaxation labeling we have considered the compatibility factor as 1.05 for optimum speed of convergence.A $4 * 4$ neighborhood was considered while updating the labels in each frame and a $3 * 3 * 3$ neighborhood was considered while updating the labels along the temporal axis. While constructing the tracks,we have considered the blocks at pixel level,$4 * 4$ block level,$8 * 8$ block level,$16 * 16$ block level and $32 * 32$ block level and found out the top N dominant tracks belonging to all these levels. For computing the cost function we have taken the value of α as 0.5 to establish the dominance of the effect of magnitude over angle of the optical flow vectors. We have already quantize the angles into 8 directions and also realigned the optical flow vectors into proper directions. So now magnitude of the optical flow vectors will have a higher stand over the direction to which it is oriented. So the magnitude is given preference in cost function over angle of orientation of the optical flow vectors.

7.3 Metrics for evaluation

Various performance measures within action classification exist, covering different aspects of the task. This section covers the most used performance measures, their benefits and drawbacks.The Craneld tests, conducted in 1960s, established the desired set of characteristics for a retrieval system. Even though there has been some debate over the years, the two desired properties that have been accepted by the research community for measurement of search effectiveness are recall,i.e., the number of action videos classified to a class; and precision, i.e., number of action videos correctly classified to a class.

7.3.1 Precision and recall

Effectiveness is purely a measure of the ability of the system to satisfy the user in terms of the proper classification of a test action. Initially, effectiveness can be measured exploiting precision and recall; a similar analysis could be given for any pair of equivalent measures. It is helpful at this point to introduce the famous confusion matrix depicted in table.

Table 7.1: Precision and Recall

Actions	Deemed non-relevant	Deemed relevant
negative	true negative (TN)	false positive (FP)
positive	false negative (FN)	true positive (TP)

Such table is a visualization tool typically used in supervised learning (where it is also called a matching matrix). Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class.

One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e., commonly mislabeling one as another).

Precision is defined as the number of correctly classified action videos by the total number of action videos (namely $\Pi = \text{TP} / (\text{TP} + \text{FP})$), and Recall is defined as the number of action videos classified to a class divided by the total number of action videos (namely $\rho = \text{TP} / (\text{TP} + \text{FN})$).

In terms of the confusion matrix above, accuracy = (TP + TN)/(TP + FP + FN + TN).

7.4 Result

We have 70 action video sequence ,where there are 10 classes and each class has 7 videos. We have employed leave one out cross validation followed by classification using Support Vector Machine(SVM). Leave-one-out cross-validation involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sampling. The SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary classifier.

Table 7.2 shows the confusion matrix created taking the 10 class,where the columns denote the actual classes and the rows denote the predicted classes. We can calculate the recall,precision of each class from the confusion matrix and thus estimate the overall accuracy of our classification system.

Table 7.3 depicts the calculated Recall,Precision,F-1 Score per class.

So,the overall accuracy of the system

$= \frac{59+56+39+52+55+49+54+51+58+47}{63*10}$

$= 82.54\%$

Table 7.2: Confusion Matrix

Actual Class Predicted Class	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10
Class 1	59	0	0	0	0	0	0	0	4	0
Class 2	3	56	0	3	0	0	0	0	1	0
Class 3	1	1	39	2	4	4	3	4	0	5
Class 4	2	2	2	52	0	0	2	1	2	1
Class 5	0	0	2	0	55	2	2	0	1	1
Class 6	1	1	2	2	1	49	2	2	1	2
Class 7	0	0	5	0	2	0	54	0	0	2
Class 8	0	0	2	0	4	0	2	51	0	4
Class 9	3	2	0	0	0	0	0	0	58	0
Class 10	2	3	1	1	2	1	3	1	2	47

Table 7.3: Precision , Reacll and F1-Score

Class	Class description	Recall	Precison	F1-Score
Class 1	Discussion of two or several people	93.65%	83.09%	88.05%
Class 2	A person gives an item to a second person	88.88%	86.15%	87.49%
Class 3	An item is picked up or put down	61.90%	73.58%	67.23%
Class 4	A person enters or leaves an office	82.53%	86.66%	84.54%
Class 5	A person tries to enter an office unsuccessfully	87.30%	80.88%	83.96%
Class 6	A person unlocks an office and then enters it	77.77%	87.5%	82.3%
Class 7	A person leaves baggage unattended	85.71%	79.41%	82.43%
Class 8	Handshaking of two people	80.95%	86.44%	83.60%
Class 9	A person types on a key-board	92.06%	85.29%	88.54%
Class 10	A person talks on a telephone	74.60%	75.80%	75.19%

Chapter 8

Future Works

There are a variety of enhancements that could be made to this system to achieve greater performance in real life situation.Each of classes of LIRIS dataset can be a normal activity, a human-human interaction or a human-object interaction, or a combination of the latter two types.

But our system fails to recognize the actions properly if there is more than one action taking place simultaneously.Our main assumption is there is only one dominant motion in the system which is depicted by the global motion of the video.But if two or more than two actions takes place simultaneously,it gives rises to more than one global motion in the system.

Bibliography

[1] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 90–102. IEEE, 1997.

[2] Claudette Cédras and Mubarak Shah. Motion-based recognition a survey. *Image and Vision Computing*, 13(2):129–155, 1995.

[3] Hae Jong Seo and Peyman Milanfar. Action recognition from one example. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):867–882, 2011.

[4] Shandong Wu, Omar Oreifej, and Mubarak Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1419–1426. IEEE, 2011.

[5] Junghye Min and Rangachar Kasturi. Activity recognition based on multiple motion trajectories. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 199–202. IEEE, 2004.

[6] Pedro Canotilho Ribeiro and José Santos-Victor. Human activity recognition from video: modeling, feature selection and classification architecture. In *Proceedings of International Workshop on Human Activity Recognition and Modelling*, pages 61–78. Citeseer, 2005.

[7] Ahmed Elgammal, David Harwood, and Larry Davis. Non-parametric model for background subtraction. In *Computer VisionECCV 2000*, pages 751–767. Springer, 2000.

[8] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2432–2439. IEEE, 2010.

[9] Min Hu, Saad Ali, and Mubarak Shah. Detecting global motion patterns in complex videos. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–5. IEEE, 2008.

[10] Ran Xu, Priyanshu Agarwal, Suren Kumar, Venkat N Krovi, and Jason J Corso. Combining skeletal pose with local motion for human activity recognition. In *Articulated Motion and Deformable Objects*, pages 114–123. Springer, 2012.

[11] Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):288–303, 2010.

[12] Mi Zhang and Alexander A Sawchuk. Motion primitive-based human activity recognition using a bag-of-features approach. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 631–640. ACM, 2012.