MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Framework for measurement of the intensity of motion activity of video segments

Kadir A. Peker
Ajay Divakaran

TR2003-64 June 2003

Abstract

We present a psychophysical and analytical framework for comparing the performance of motion activity measures for video segments, with respect to a subjective ground truth. We first obtain a ground truth for the motion activity by conducting a psychophysical experiment. Then we present several low-complexity motion activity descriptors computed from compressed domain block motion vectors. In the first analysis, we quantize the descriptors and show that they perform well against the ground truth. The MPEG-7 motion activity descriptor is also among the best performers. In the second analysis, we examine the specific cases where each descriptor fails, using a novel pair-wise comparison method. The analytical measures overestimate or underestimate the intensity of motion activity under strong camera motion or extreme camera angles. We finally discuss the experimental methodology and analysis methods we used, and possible alternatives. We review the applications of motion activity and how our results relate to them.

Publication History:–

1. First printing, TR2003-64, June 2003

Framework for measurement of the intensity of motion activity of video segments

Kadir A. Peker, Ajay Divakaran

Mitsubishi Electric Research Laboratories,

201 Broadway, Cambridge, MA 02139, USA

Tel: +1 (617) 621 7500. Fax: +1 (617) 621 7550.

{peker, ajayd}@merl.com

ABSTRACT

We present a psychophysical and analytical framework for comparing the performance of motion activity measures for video segments, with respect to a subjective ground truth. We first obtain a ground truth for the motion activity by conducting a psychophysical experiment. Then we present several low-complexity motion activity descriptors computed from compressed domain block motion vectors. In the first analysis, we quantize the descriptors and show that they perform well against the ground truth. The MPEG-7 motion activity descriptor is also among the best performers. In the second analysis, we examine the specific cases where each descriptor fails, using a novel pair-wise comparison method. The analytical measures overestimate or underestimate the intensity of motion activity under strong camera motion or extreme camera angles. We finally discuss the experimental methodology and analysis methods we used, and possible alternatives. We review the applications of motion activity and how our results relate to them.

Keywords: Motion Activity, Compressed Domain Feature Extraction, Video Indexing, MPEG-7

1. INTRODUCTION

Indexing of the vast amounts of digital video content for browsing, retrieval and summarization purposes has been an active research topic in recent years. The MPEG-7 standard for the description of multimedia metadata is developed to cover key technologies in this area. It covers a number of features, or descriptors (Ds), of video content such as shape, color, motion, etc. used for indexing. Motion activity is one of the descriptors included in the MPEG-7 specification [1][2].

1. Motion Activity

The intensity of motion activity is a subjective measure of the perceived intensity, or amount, of motion activity in a video segment. A talking head in an interview is usually a low activity segment, whereas a close-up shot of a slam-dunk in basketball is perceived as high activity. Note that it is different from camera or global motion in that it considers the overall perceived intensity of motion activity in the scene.

A number of low-complexity measures have been used to describe such motion activity characteristics of video segments [1][3][4][5][6][7][8]. Pfeiffer et al use a combination of image and audio features to determine the activity level of video segments and use that information in selecting interesting segments from video for summarization [9]. Vasconcelos et al use a ‘tangent distance’ between consecutive frames as the measure of motion activity, and use it to characterize video sequences in terms of their action or romance content [10]. Wolf uses the mode of motion vector magnitudes as the measure of activity level, which he then uses to find the most still image in a video segment, and selects it as a representative frame [8]. The motion activity descriptor used in MPEG-7 is the variance of the motion vector magnitudes, which is readily computable from MPEG compressed video streams [1]. Motion activity is interpreted as a measure of “summarizability” or the entropy of a video segment in [5] and [6], and based on this interpretation, the average of magnitudes of block motion vectors is used in summarizing video segments. A similar motion activity descriptor is used in detecting interesting events in sports video [11][12].

The motion activity descriptor enables applications such as video browsing, surveillance, video content re-purposing and content based querying of video databases [13]. It is more effective when the video content consists of semantic units

that significantly differ in their motion activity levels. For instance, in a news video, anchorperson shots are very low activity, whereas outdoors footage or the sports segments have higher activity levels [5][6]. Still, motion activity has its strength in the very low-cost, compressed domain descriptors for it, which allow very fast pre-filtering of data, or dynamic and interactive browsing and summarization applications where the user, or further automatic processing tolerates the initial low precision. We believe that the machine vision techniques, which mostly involve optical flow and segmentation, are not easily applicable to natural video for now, especially within the constraints of the applications where the data volume is large and higher speeds that will allow interactivity is desired.

2. Performance evaluation of motion activity measures

There are two possible approaches to the performance evaluation of a motion activity measure. One is as an estimator of the perceived subjective motion activity as described in the introduction. The second is as an analytical measure that is used within a specific application context, where it is considered successful to the degree that it contributes to the performance of the overall application. In the second case, the conformance of the descriptor to the perceived intensity of motion activity is not of primary concern.

In the case of MPEG-7, the motion activity descriptor is defined as an estimator of the subjective motion activity. Hence, an evaluation of the alternative motion activity measures using a subjective ground truth is necessary. The ground truth in this case is the perceived intensity of motion activity in a given video segment evaluated by human subjects. While the MPEG-7 descriptor has been developed with a ground truth data-set of 622 video segments [14][15], the data-set lacks statistical data about the subjects and also has segments that vary in both length and quality of shot segmentation. This has made it difficult to assess the efficacy of any automatically computed descriptor of the intensity of motion activity. In [16], [17], and [18], we provide a psychophysically sound basis for subjective and objective measurement of the intensity of motion activity.

In the following sections, we describe the psychophysical framework for the measurement of the intensity of motion activity of video segments, and the evaluation of the performance of different analytical measures of motion activity. First, we construct a test-set of video segments carefully selected so as to cover a wide variety and dynamic range of motion activity. We conduct a psychophysical experiment with 15 subjects to obtain a ground truth for the motion

activity. Then we present several low complexity motion activity descriptors computed from MPEG motion vectors in the compressed domain.

We compare the motion activity of the video segments in the test set as assessed by the subjects, and as computed by described analytical descriptors. In the first comparison method, we quantize the analytical descriptors and compute the error with respect to a ground truth computed across the subjects. In the second comparison method, we compare the test video segments in pairs to determine the pairs where one video segment is unanimously rated as higher activity than the other by the subjects. Then for each analytical descriptor, we find the number of such pairs where the descriptor fails to give the correct ordering. Based on these results, we examine the specific cases (pairs of video segments) where each analytical descriptor, and motion vector based descriptors in general, tend to fail. We verify our initial subjective observation that the distance from camera, and strong camera motion are main cases where motion vector based descriptors tend to overestimate or underestimate the intensity of motion activity. We also show that the variance of the magnitude of motion vectors, on which the MPEG-7 motion activity descriptor is based, is one of the best among the descriptors tested.

2. THE GROUND TRUTH FOR MOTION ACTIVITY

1. Selection of the Test Clips

We select 294 video segments of length 1.5 seconds from the whole MPEG-7 test content. The number and duration of clips are chosen as a compromise between viewer fatigue, memory effects, etc. and sufficient data size for analysis, sufficient duration for perception, etc. We select the test clips from over 11 hours of MPEG-7 test video sequences through several elimination steps in order to cover a diverse range of semantic activity types and activity levels. We first use biased random sampling to have a better distribution of motion activity levels in the test set, and then manually reduce the set to 294 to cover a variety of cases as listed in Table 1.

2. The Psychophysical Experiment

We used 15 subjects in our experiment. The subjects were naïve as to the purpose of the experiment so as to obtain unbiased results. Subjects were shown what to expect in terms of the minimum and the maximum level of activity to be displayed so that they maintain a consistent mental scale through out the experiment. Figure 1 shows samples from the

training set the subjects were shown to illustrate the extent of motion activity levels. However, they were not given a complete scale, i.e. a specification of what each activity level should cover. Motion activity is very subjective and it is not possible to objectively determine what activity levels are to be labeled 0, 1, 2, etc. Hence, once the maximum and the minimum available activity levels were shown, the subjective activity scale was left up to the subjects.

We arbitrarily selected 6 activity levels, from 0 (lowest) to 5 (highest). This is a number that strikes a balance between two conflicting requirements: There must be a small number of levels so that humans can easily classify clips. On the other hand, we need a large number of levels so as to correlate automatically derived measures of activity to those obtained by humans with an adequate resolution.

Test video clips were displayed to the subjects consecutively, in random order. The subjects were asked to play each video clip only once and decide based on the perception of motion activity they got at first, and try not to be affected by other factors such as semantics.

3. Results of the Psychophysical Experiment

We verify the consistency of the subjective results and examine how well subjects agree on the motion activity of the test clips, as a first step. This analysis provides us with an understanding of the precision of the subjective ground truth, and enables us to determine an acceptable range of error between the ground truth and the computed descriptors.

We first compute the median of the subjects' activity evaluations for each clip as a ground truth. Then we compute each subjects deviation from this ground truth (the sum of absolute differences). Table 2 shows the total and average (per video clip) deviation of each subject. The median deviation of the subjects with respect to the ground truth is 0.73. We will use this figure as a guideline in evaluating the performance of the analytical measures later. Note that the error values suggest that subject 14 is an outlier. We also find that the majority of the subjects agree on either of two activity levels for each clip. These results show that there is reasonable agreement among subjects on the subjective intensity of motion activity.

3. MOTION VECTOR BASED MEASURES OF MOTION ACTIVITY

We use a number of low-complexity measures computed from compressed domain MPEG block motion vectors. Although the compressed domain motion vectors are not precise enough for object motion analysis, they are sufficient for the measurement of the gross motion in video. The low-complexity and compressed domain computation of the measures allow low-cost, high speed processing of large amount of video data, allowing applications such as data reduction through pre-filtering, real-time processing, and dynamic interactive summaries. It also enables the implementation of these measures on consumer device platforms within practical cost and complexity limits.

The most trivial measure of the overall motion activity in the video that we use is the average of motion vector magnitudes (*avg*), which assumes that the faster the objects move the higher the perceived activity is. The median of the magnitudes (*med*) is another option, which is expected to be more effective in filtering out the spurious motion vectors common in motion compensated video.

A few empirical observations and hypotheses lead to more measures: The first observation is that the perceived motion activity is higher when the motion is not uniform. Uniform or homogenous motion is perceived as a lower activity than predicted by the average of motion vector magnitudes. A typical example is camera motion where the average of motion vector magnitudes is very large but the perceived activity is not as high as it would be for an activity due to object motion. The variance of the motion vector magnitudes (*var)* is a descriptor that is motivated by this observation. The MPEG-7 motion vector descriptor is also based on the variance of motion vector magnitudes. The *mean0* and *mean1* descriptors are aimed at removing spatial uniformity of the motion vector field. They are computed by first subtracting the average of motion vectors in a frame, and then computing the magnitudes of the vectors and averaging. This operation, for instance, would eliminate a significant portion of a camera pan. *mean1* is different from *mean0* in that, the frame is divided into 12 blocks and the vector average is computed for each block separately. The *diff* descriptor, which is based on temporal differentiation of the motion vector field, is aimed at removing the temporal uniformity. Please see [19] for a detailed description of these measures.

Another empirical hypothesis or observation is that the perceived activity can be high even when there is a relatively small but fast moving object in a low activity background, where the average of motion vector magnitudes is low. We

use the maximum of the motion vector magnitudes (*max*) in our experiments to test this hypothesis. We also use two variations of the maximum – *max1* and *max2*, which discard top 1.5% and 10%, respectively, of motion vectors sorted by magnitude, in order to filter out spurious vectors or very small objects.

We compute the above nine measures (descriptors) for each P frame of the mpeg coded video segments. Intra-coded macroblocks, which do not have a coded motion vector, are assigned a zero motion vector. We compute the motion activity of each video segment by averaging the P-frame descriptors. Thus, we compute nine descriptors for each video segment. Another set of tests conducted using the maximum over frames in a segment, instead of the average, also gave similar results to the ones presented here.

4. AVERAGE ERROR PERFORMANCE OF THE DESCRIPTORS

The first analysis of performance of the descriptors is based on the average error with respect to the ground truth over the whole data set [18]. We use the median of the subjects' evaluations as our ground truth so as to minimize the effect of outliers. Taking the mean of the subjective levels would assume a linear scale for the subjective motion activity, which is not necessarily true. Rounding of the mean, as well, is problematic in the context of discrete activity levels of 0 to 5.

We quantize each descriptor into activity levels 0 to 5. We find the quantization thresholds for each descriptor by minimizing the error between the ground truth and the quantized descriptor. In order to separate the training set and the test set, we use half of the 294 clips to optimize the quantization thresholds, and conduct performance evaluations on the other half. We split the data set of 294 clips into two sub-sets randomly each time, and repeat this process 30 times. We use the sum of absolute differences between the quantized descriptor values and the ground truth as the measure of error.

3.1. The Quantization of the Descriptors

We quantize the descriptors into levels 0 to 5 so that they can be compared to the ground truth. The scatter plot of descriptors vs. the ground truth in Figure 2 suggests that the relationship between the descriptors and the ground truth is not clear enough to fit a curve. We formulate the problem as optimal quantization of the descriptor values to minimize the error between the quantized values and the ground truth. Note that this is not a quantization problem in the conventional sense where the error is between the unquantized and the quantized values. In the common quantization

problem, the original set of unquantized values alone determines the quantization. In our case, the optimal quantization is determined by the second set of values, which is the ground truth.

The following is a formulation of the optimal quantization problem (see Figure 3). The set of b_i that minimizes *err* is the optimal solution.

$$q_i = \begin{cases} 1, \text{ if } 0 < d_i < b_1 \\ 2, \text{ if } b_1 < d_i < b_2 \\ \ldots \end{cases} \quad \text{, where:} \quad \begin{array}{ll} s_i & \text{Subjective activity for clip i} \\ d_i & \text{Descriptor value for clip i} \\ q_i & \text{Quantized descriptor value for clip i} \\ b_1..b_5 & \text{Quantization boundaries} \end{array}$$

$$err = \sum_{i=1}^{294} | s_i - q_i | \qquad (1)$$

We optimize each threshold value separately, independent of the others. We first consider the first threshold b_1. In our case of 147 descriptor values in the training set, we quantize all the descriptors up to the nth to value 0, and all the others to value 1, where n is variable. We compute the error for each value of n=0 to 148 (including the positions before the first and after the last descriptor values). In [16], we show that the optimal value for b_1 is the one that minimizes the error. The other four threshold values are computed in the same way.

The following formulation shows how we find b_1, the first quantization threshold. We select the b_1 that minimizes *err*.

$$b_1: \qquad q_i = \begin{cases} 0, \text{ if } 0 < d_i < b_1 \\ 1, \text{ if } b_1 < d_i \end{cases}$$

$$err = \sum_{i=1}^{294} | s_i - q_i | \qquad (2)$$

$$\ldots$$

4. Average Error Results for the Descriptors

We provide the average error results for each descriptor in Table 3. The first observation is that all the descriptors perform acceptably, considering the median subject error of 0.73. The descriptors *max1*, *max2*, and *var* – which is essentially the descriptor used in MPEG-7 – are the best performers according to the average error results. The refinements made to the *avg* descriptor through the *mean0* and *mean1* descriptors improve the average error

performance. However, the approach of *diff* does not seem to provide an improvement in estimating the subjective activity level. The *median* descriptor also performs inferior to the *avg*. We observed that the median of motion vector magnitudes can be very low, sometimes 0, if the activity in the frame covers a small area.

5. Comparison of Subjects' and Descriptors' Errors

We note that in some cases the descriptors perform better then some of the subjects, although the descriptors are estimators of the subjects' evaluations. The reason why the descriptors perform so well compared to the subjects can be the bias of individual subjects towards high activity or low activity. In other words, the reason can be the different mental scales of activity that each subject has. A slight shift in the subjective evaluations of a subject can add up to a significant error in the total. The quantization of the automatically computed descriptors, by definition, aligns them well with the ground truth. Figure 4 illustrates how the subjective errors are biased either towards +1 or –1 for almost each subject. The problem of quantization of the descriptors and the mismatch between subjective scales motivate us for the pairwise comparison method that we describe next.

5. PERFORMANCE ANALYSIS USING PAIRWISE COMPARISONS

1. Limitations of the Average Error Analysis

The average error analysis described in the previous section validates the proposed descriptors as acceptable estimators of the subjective level of activity, and shows their comparative performance. However, the aforementioned framework of analysis based on derivation of a ground truth from 15 subjective evaluations and quantization of the computed descriptors does not allow for a more precise and detailed performance study. Note that we need to make certain assumptions in order to overcome a number of issues before we can obtain a unique subjective activity value and a quantized descriptor for each clip. More specifically, we can identify the following issues:

- The fact that the subjective scale of activity varies from subject to subject (do they mean the same thing when they say *activity level 2*?).
- The issue of selecting a single activity value when the subjects do not agree.
- The issue of comparing the continuous descriptors to the discrete 6 level ground truth.

All these non-idealities limit us to using averaged error measures over the whole data set to compare the various descriptors, rather than examine their validity and performance on the level of individual clips. At the individual clip level, it is hard to assess the performance of a descriptor by looking at the difference between the ground-truth absolute activity level and the quantized computed descriptor. Through our subjective observations, we know that the computed descriptors sometimes give counter-intuitive results in certain cases, even though they perform as well as human subjects in terms of average error. For example, most of the motion vector based descriptors consistently overestimate activity level when there is strong camera motion or when a moving object is in a close-up. We want to investigate these limitations of the descriptors. Hence, we propose a new framework that enables a more detailed analysis to observe the descriptor performance on individual clips [16][17].

2. The Pairwise Comparison Method

We define a transitive 'greater than' relationship between video segments in terms of their motion activity level. We consider whether the subjects agree on one video segment having higher activity than the other, rather than on the absolute activity level of those clips. For instance, let us assume that subjects A and B evaluate the video clip *i* as having activity level 2 and 4, respectively. If they evaluate clip *j* as activity level 3 and 5, respectively, then they both agree that clip *j* has a higher activity level than clip *i*. Then we can test if the computed descriptors agree with this assessment. In this way, we do not need to resolve the conflict between different subjective activity scales, and we do not need to devise a mapping from the continuous descriptor space to the discrete subjective activity levels. The statement that clip *i* has a higher activity level than clip *j* is more reliable than the statement that clip i has activity level 3.

It is well known that humans are much better at making comparative judgements such as "clip A has a higher activity level than clip B," than at judging the absolute activity levels of clip A and clip B. But a pairwise comparison of 294 clips would require (294x293)/2=43,071 video pairs to be viewed, which is infeasible with human subjects. Even an experiment with 4000 pairs is not feasible. Hence, the pairwise method described here provides a practical way of achieving similar results.

In our analysis, we define the 'greater than' relationship as follows:

Definition: *The activity level of clip i is greater than the activity level of clip j if and only if all the subjects assign a higher activity level to clip i than they assign to clip j.*

We require unanimity to guarantee transitivity, and to minimize human error. We obtain 4134 such pairs, i.e. for 4134 pairs in our data set, one clip is unanimously rated as higher activity than the other clip. Since the relationship is not defined for all the pairs, we do not have a complete ordering of the 294 clips. Rather, we have a number of ordered lists.

To further understand the topology of the resulting relationship structure, we model the relationship as a directed graph as in Figure 5. Note that we have a three-level structure, and a set of eight clips that have no relationships to the rest. Some of these eight clips have conflicting semantic context and activity, such as a slow replay of a goal-scoring moment.

3. Pairwise Comparison Error Results

We have a set of 4134 ordered pairs as described above. We find the number of pairs where the computed descriptors order the pairs in the opposite way that the human subjects order them. Table 4 shows the total number of such pairs, i.e. the error figure, for each of the descriptors.

The descriptor *max2* makes the fewest errors at about 5%. The rest of the errors are between 7% and 10%, except for the *max0*, *max1* and the *diff* descriptors. Hence, the descriptors in general give a correct result for more than 90% of the 4134 pairs. The *max2* descriptor has the lowest error value in both the average error analysis and in the current analysis. Also the descriptors *var* and *mean0* perform well in this analysis as they did in the average analysis. *max1* and *max0*, in contrast, give poor results in this analysis although they had low average errors.

We group the descriptors into three according to their performance in this error analysis. *max2*, *var* and *mean0* are the best performing descriptors in the pairwise comparison analysis, as well as the average error analysis in. They have errors ranging form 5% to 8%. The *median* descriptor also performs well in the pairwise test. *mean1* and *avg* are in the second group with 9% to 10% error. *diff*, *max1* and *max0* are in the last group with errors over 10%. This general grouping is in accord with the average error analysis except for the *max0* and *max1* descriptors.

4. Individual Characteristics of the Descriptors

We examine the pairs of video clips where each descriptor fails, in order to gain insight into the individual characteristics of the descriptors. Firstly, for 75 of the 4134 pairs, none of the descriptors gives the correct ordering. This set of 75 pairs where all the descriptors fail gives us an idea about the general fallacies of motion vector based descriptors. The pairs in this set mainly fall into two categories: 1) A strong camera motion or an object in close-up that results in large motion vectors, but not perceived as high activity. 2) Activities such as dancing figures, sports or light effects that are perceived as high activity but do not result in large motion vectors because of a wide camera angle or small object size (See Figure 6 for samples). Some of the descriptors are devised to overcome these effects, such as *mean0-1*, *var*, *diff* and *max0-2*. In fact, they improve upon the trivial descriptors such as *avg*. However, there still are cases where the heuristics do not suffice and a semantic understanding of the content is necessary. Note that the two misleading factors mentioned above not only affect motion vector based descriptors, but any descriptor based on the change between consecutive frames.

We examine the correlation between the descriptors by finding the number of common video clip pairs for which they fail. We find that *var* has highest correlation with *mean0*, *max1* and *max2*. *max2* also has highest correlation with *var* and *mean0*. Thus, the best performing three descriptors have similar characteristics. Note that *var* and *mean0* are mathematically similar as well. *diff* has high correlation with *var*, *mean0* and *max1*, hinting at that it performs similarly in problematic cases such as camera motion. The *median* and the *max0* have little correlation with other descriptors. This suggests that *median* can be a candidate to complement the other descriptors in certain cases.

6. DISCUSSION ON PSYCHOPHYSICAL EXPERIMENTS AND STATISTICAL ANALYSIS

1. Selection of the data set

The randomness of the test set is an important factor in statistical analysis of experiments. Any such psychophysical experimental test set should not be biased towards a particular subset where some of the measures are particularly strong or weak. In our case, the test set should provide a fair representation of the “universe” of video segments that we want the motion activity descriptor to be applied to. We could not carry out a simple random sampling of the initial MPEG-7 test set because it does not fully reflect the variety of content in the set of all existing video sequences. Hence, we first

carried out a biased random sampling to have a more uniform distribution of motion activity levels. Second, we hand picked about 300 segments out of 5000, to represent a good variety of motion activity types and levels.

Another important consideration in selecting the test set is the target use of the tested measures. Since, in the context of MPEG-7, we were testing the available measures for generic use and without an application constraint, we tried to cover as wide a range as possible. For a specific application such as news video browsing, a more targeted experiment could be designed, which would compare the performance of measures for that purpose.

The other extreme of a completely random selection of test clips would be a very controlled set of test clips. This approach can be used to test a hypothesis under controlled experiments. For instance, shots of speeding cars at different speeds can be used to test the effect of speed, or a set of crowd shots with different number of people can be used to test the effect of object number on motion activity perception. Even further would be the use of synthetic stimuli consisting of a controlled configuration of objects/shapes, velocities, trajectories, etc. This would be similar to the methodology used psychophysical experiments. It would help to understand the basics of the phenomenon, but would have limited direct application to natural video. In this paper our focus is on experiments that have a direct impact on applications.

2. Considerations in designing the experiment

Selection of the subjects

A sufficient size of data is necessary to obtain reliable statistical results. However the size is limited by practical constraints such as the number of accessible subjects, human fatigue, time limitations, etc. Expert help is valuable in designing these aspects of the experiment. The subjects are also required to be naïve to the purpose of the experiment in order to avoid any bias.

Preparation of the subjects

Calibration of the subjects is an important step in psychophysical experiments. A warm-up set helps the subjects to construct a consistent mental scale and maintain it through out the real experiment.

An important decision in our experiment was not to give an a priori definition of the "intensity of motion activity" to the subjects, but rather give a vague definition in terms of examples of the extreme cases. We left the exact definition, the scale, and specific criteria of the intensity of motion activity to the subjects. In the end, the ground truth that we obtained is itself a definition of what subjective intensity of motion activity is. The reason for this decision is that the MPEG-7 motion activity descriptor is a generic descriptor and is not tied to a specific application context. If there were an application at hand, we would have given subjects more specific guidelines in assessing the intensity of motion activity of video segments, such as what should be considered more activity and what should be less. The subjects would have had better agreement in that case. Still, the agreement level among subjects was satisfactory in our results.

We instructed the subjects to ignore the semantic connotations of the video content and directly concentrate on the perceived intensity of motion activity. However, it is virtually impossible to isolate the effects of the semantics completely. As a matter of fact, we observed the influence of semantics when examining the failures of analytical measures in pairwise analysis. The designer of the experiment has to decide how strongly the semantic effects should be minimized, or if they should directly be part of the picture.

Other experimental parameters

In many applications, we observe that a higher resolution of motion activity than six levels is necessary, and is possible within the constraints of a domain. Another experimental method that will allow higher resolution is side-by-side pairwise comparison of video clips by the subjects. The human visual system is much more successful in comparing two stimuli at hand than assessing the absolute intensity of stimuli. However, in such an experiment, the data size would have to be much smaller because of the inherent combinatorial expansion.

3. Statistical analysis

A variety of statistical tools are applicable once we collect the data. We chose to quantize the analytical measures, take the median across the subjects, and then compute the error between the two. We also tried curve fitting and regression analysis. Linear regression was not suitable for the data. Non-linear curve fitting also didn't provide useful solutions without going into under or over-fitting. By also incorporating our domain knowledge, we formulated the quantization

problem as a monotonic mapping to the discrete range of subjective motion activity levels. To avoid the circular problem of training and testing on the same set, we used cross-validation by dividing the set into two.

We also tried various clusterings of the data: Clustering subjects according to the way they rate test clips, clustering test clips using subjects, clustering test clips using analytical measures, etc. Using hierarchical clustering, we looked for subsets of the test set where certain clusters of subjects agree, etc. This was targeted at understanding the factors in perception of motion activity and how different subjects were influenced by different factors. Our preliminary investigation was inconclusive, but it suggested that with more elaborate and specific experiments we could gain more insight into how motion activity is perceived for different categories of content. Because of the limited size of the data set we had, we were cautious about investigating categories by further dividing the data set into subsets.

Further statistical analysis of the data is possible by tabulating a number of factors/measures such as camera motion, object distance, or even the number of or coverage of objects, average speed, semantics (sports, etc). Then we can see how these factors relate to the motion activity of the clips determined by the subjects and computed by analytical measures.

In short, we can view it as a data mining problem where we look at all the data from subjects, analytical measures, other information on clips, etc., and try to observe patterns, principal components of motion activity and its subjective perception. However such an approach would be time consuming. Therefore, we limited the scope of our analysis using our knowledge of the domain and were able to get useful insight.

4. The pairwise comparison method

The pairwise comparison method provides a partial ordering of the test set. This novel method solves many problems inherent in the average analysis method we used first. It provides an effective framework for evaluating the performances of the analytical measures in a more detailed and precise way. However, it is limited to the subset of pair of test clips where there is an ordering based on unanimous agreement of the subjects. Since this subset is where the relationship is most “obvious”, the pairwise method in a sense catches the worst errors. In that way, average analysis still has its place in understanding the overall performance of the tested measures.

Pairwise comparison of test data points of subjective data, in general, is an effective way of obtaining a higher precision subjective information from a lower precision set of subjective data. Note that we define the relationship between two video clips in terms of the relative assessment of each subject and not their absolute assessments. That is, we consider whether all the subjects agree that one of the clips is higher-activity than the other one. It is in contrast to considering the video clips where all the subjects agree on a single absolute activity level. The latter is still susceptible to the problems due to the variations in the mental scales of the subjects, the limitations of the human visual system on assessing absolute intensity of stimuli, etc. And we would still need to quantize the analytical measures to compare to the absolute subjective activity levels. The pairwise comparison method bypasses these problems.

There are a number of improvements or extensions that we can think of the pairwise method we used in this experiment. One possibility would be to relax the 'greater than' relationship such that it holds for a larger subset of the pairs, yet maintains the transitive and reflexive property – hence providing a partial ordering of the set. A further extension could be to define a confidence measure based on the degree of agreement among subjects, and then scale the error for each pair with the confidence measure.

5. Motion activity as tested, and as used in applications

Lastly, we want to look at some of the applications where motion activity descriptors are used. We note that motion activity is a gross descriptor of the motion activity feature. Hence, it is best used for pre-filtering of data before more precise and costly operations are carried out. It can also be used to complement other low-level audio-visual descriptors. The low complexity of the motion activity descriptors we presented here enables interactive applications. In such applications, since the user can quickly recover from errors, s/he can more than compensate for the moderate precision.

The results we show in this paper establishes that simple measures based on compressed domain block motion vectors reasonably estimate the perceived motion activity. Thus, we are able to confidently use such measures in various applications. For instance, we interpret the motion activity of a video segment as a measure of its "summarizability" in [5] and [6]. Based on this interpretation, we determine the number of keyframes necessary to summarize a video segment. In that work, we rely on the established validity of simple measures of motion activity. In [16] and [20], we

employ a similar “entropy” interpretation of motion activity of a video segment to vary the playback speed of video segments as to maintain a constant pace throughout a video sequence. Again, the intuition is based on the subjective notion of motion activity, and we rely on the results presented in this work to use simple motion activity descriptors. Thus, even if an application uses a motion activity measure without caring about its conformance to the perceived motion activity, this study provides the foundation that the simple measures of motion activity are in fact meaningful and valid descriptors.

The experimental results we presented here not only establish the validity of the motion activity descriptors, but also describe the non-linear relationship between the analytical descriptors and the perceived motion activity. As an application of these results, the quantization of the descriptors developed here is used in [6] to make the problem tractable.

The next logical step in this line of subjective experimentation would be to design psychophysical experiments designed directly for specific applications. For example, for the variable playback speed application, we could ask the subjects to increase the playback speed of video segments until they can no longer follow the action. This way, we would get a measure similar to motion activity, but directly tuned for the application at hand. Another possibility is to limit the video content to a specific domain such as news or sports, and use side-by-side comparison by the subjects to achieve higher precision and resolution in measuring motion activity.

7. CONCLUSIONS

We reviewed our past work in which we established a subjective ground truth for motion activity and an analysis framework for the performance of descriptors. We showed that the low-complexity, motion vector based descriptors proposed in this paper are acceptable estimators of motion activity. We analyzed the descriptors’ performance and how they compare to each other in terms of average error. We then presented a novel pairwise comparison analysis method to investigate the performance of the descriptors in detail. We showed that the variance, which is a part of the MPEG-7 standard, is one of the best performing descriptors along with *max2* and *mean0*, which are novel descriptors. The average of motion vectors, which is a simple way of estimating activity, commonly used by many researchers, provides an acceptable estimation of motion activity as well.

Our review establishes that the design of the experiment is valid and reasonable, by thoroughly examining the underlying issues and design alternatives. We then discussed the significance of the results and their impact on applications of the motion activity descriptor. We found that the experiments provide a solid basis for application of common measures of motion activity to different tasks such as video summarization.

REFERENCES

[1] "MPEG-7 Visual part of the XM 4.0," ISO/IEC MPEG99/W3068, Maui, USA, Dec. 99.

[2] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," Proc. IEEE Trans. Circuits and Systems for Video Technology, vol. 11, no. 6, June 2001

[3] A. Akutsu, Y. Tonomura, H. Hashimoto, and Y. Ohba, "Video indexing using motion vectors," Proc. SPIE Conference on Visual Communications and Image Processing, SPIE Vol. 1818, pp. 1522-1530, 1992.

[4] E. Ardizzone, M. La Cascia, A. Avanzato, and A. Bruna, "Video indexing using MPEG motion compensation vectors," Proc. IEEE International Conference on Multimedia Computing and Systems, 1999.

[5] A. Divakaran and K. A. Peker, "Video summarization using motion descriptors," Proc. SPIE Conf. on Storage and Retrieval from Multimedia Databases, San Jose, CA, Jan 2001.

[6] A. Divakaran, R. Regunathan, and K. A. Peker, "Video summarization with motion descriptors," Journal of Electronic Imaging, vol. 10, no. 4, Oct. 2001

[7] V. Kobla, D. Doermann, K.-I. Lin, and C. Faloutsous, "Compressed domain video indexing techniques using DCT and motion vector information in MPEG video," Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases V, SPIE Vol. 3022, pp. 200-211, 1997.

[8] W. Wolf, "Key frame selection by motion analysis," Proc. of ICASSP 96, Vol. II, pp. 1228-1231, 1996.

[9] S. Pfeifer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting digital movies automatically," J. Visual Comm. Image Representation, vol. 7, no. 4, pp. 345-353, Dec 1996.

[10] N. Vasconcelos, A. Lippman, "Towards semantically meaningful feature spaces for the characterization of video content," Proc. of ICIP97, 1997.

[11] K.A. Peker, R. Cabasson, and A. Divakaran, "Rapid generation of sports video highlights using the MPEG-7 motion activity descriptor," Proc. SPIE Conf. on Storage and Retrieval for Multimedia Databases, San Jose, CA. Jan 2002.

[12] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with Hidden Markov Models," Proc. ICASSP, IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP), Orlando, FL, May 2002.

[13] D. Manoranjan, A. Divakaran, B. S. Manjunath, Ganesh R, and V. Vinod, "Requirements for an activity feature and descriptor in MPEG-7," ISO/IEC JTC1/SC29/WG11/MPEG99/M4485, Seoul, Korea, March 1999.

[14] A. Divakaran, K. A. Peker, H. Sun, and A. Vetro, "A supplementary ground-truth dataset for intensity of motion activity," ISO/IEC MPEG00/m5717, Noordwijkerhout, Netherlands, March 2000.

[15] A. Divakaran, H. Sun, H. Kim, C. S. Park, B. S. Manjunath, X. Sun, H. Shin, V.V. Vinod, & al., "Report on the MPEG-7 core experiment on the Motion Activity Feature," ISO/IEC MPEG99/m5030, Melbourne, Australia, Oct. 2000.

[16] K.A. Peker, "Video indexing and summarization using motion activity," Ph.D. Dissertation, NJIT, 2001.

[17] K. A. Peker and A. Divakaran, "A novel pair-wise comparison based analytical framework for automatic measurement of intensity of motion activity of video," Proc. IEEE Int'l Conf on Multimedia and Expo ICME 2001, Tokyo, Japan, Aug. 2001.

[18] K. A. Peker, A. Divakaran, and T. V. Papathomas, "Automatic measurement of intensity of motion activity of video segments," Proc. SPIE Conf. on Storage and Retrieval from Multimedia Databases, San Jose, CA, Jan 2001.

[19] K. A. Peker, A. A. Alatan and A. N. Akansu, "Low-level motion activity features for semantic characterization of video," Proc. of IEEE International Conference on Multimedia and Expo 2000.

[20] K. A. Peker, A. Divakaran and H. Sun, "Constant pace skimming and temporal sub-sampling of video using motion activity," Proc. IEEE Int'l Conf. on Image Processing, Thessaloniki, Greece Oct. 2001.

Table 1. A number of activity types used as a guide in constructing our test clip-set. We roughly classify activity levels as low, moderate and high. Based on our previous experience with motion activity and its measurement, we also define a fourth category where the motion vectors are very high but the perceived activity level is not proportionately high.

High Activity Content:

Music, Dance

One – few – crowd,

Close-up – medium – wide,

Lighting changes, various camera effects,

High activity – medium activity

Sports

Wide angle – close-up,

Fast – slow,

Basketball – Soccer – golf etc, Runners – bicycle race,

Score – attack, running, jumping in soccer

Effects

Logo,

Fire, explosion, etc.

Moderate Activity Content:

Drama, Education, News, etc.

Standing or still people – Walking – Running,

One – few – many…

Misleading High Activity Content:

Camera Operation (Pan, Zoom, etc…)

On static or very low activity – low or moderate activity object,

High activity background – still background,

Steady camera – unsteady camera,

Indoors – outdoors, etc.

Close-ups

Slow-moving object – fast-moving object,

Short duration and fast activity (sudden gestures)

Low Activity Content:

Talking Heads

Anchor – interview,

Still background – low motion. background (set or outdoors) – high motion background, outdoors, in crowd, etc.,

One – few.

Surveillance

People – speedway,

One – few – many,

Slow – fast – still ,

Close – far,

Also: Cartoons, commercials, scene cuts and fades

Table 2. Each subject's deviation (Sum of Absolute Differences over 294 clips) from the ground truth. The ground truth is defined as the median of the subjects for each clip. Note that the median of the subjects' deviations is 0.73.

Subject No:	10	13	7	2	12	6	3	5	8	1	4	15	9	11	14
Subjects total deviation from the ground truth	165	167	169	178	182	198	209	214	228	235	240	263	271	271	375
Subjects average deviation (per test-clip)	.56	.57	.57	.61	.62	.67	.71	.73	.77	.80	.82	.89	.92	.92	1.27

Table 3: Average error (average of absolute differences between the quantized descriptor and the ground truth) for 9 descriptors.

	max1	max2	var	max0	mean0	mean1	avg	median	diff
Average error	0.730	0.743	0.746	0.746	0.781	0.792	0.816	0.824	0.826

Table 4. Number of pairs out of 4134, for which the descriptors fail are shown, along with percentages.

	max1	max2	var	max0	mean0	mean1	avg	median	diff
Number of errors	494	218	318	827	318	389	416	337	501
Percentage	11.9%	5.3%	7.7%	20.0%	7.7%	9.4%	10.1%	8.2%	12.1%

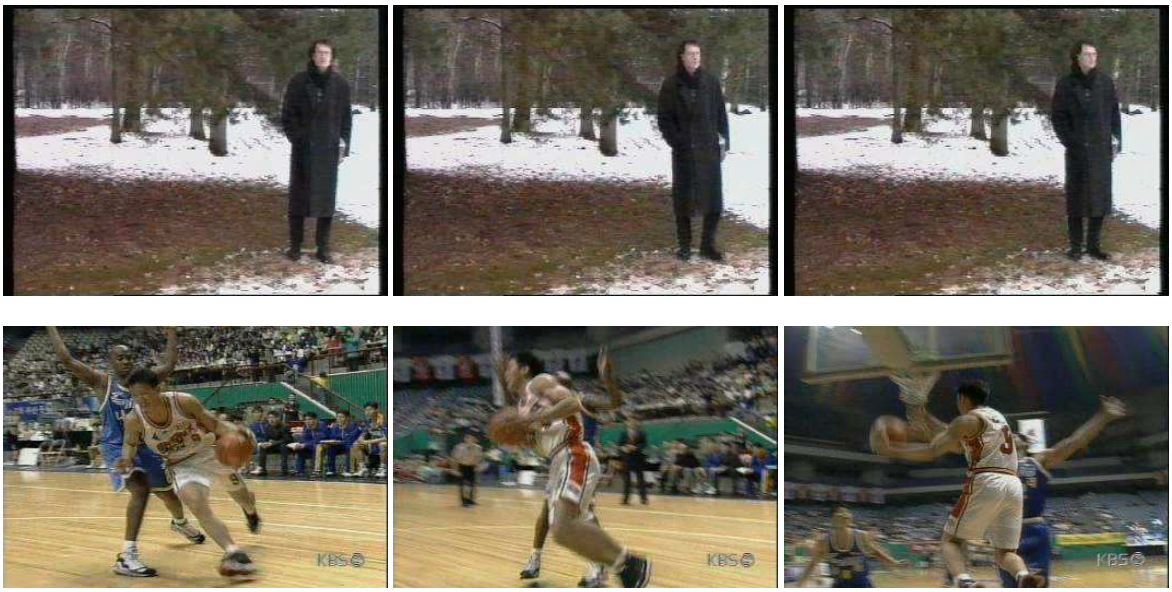


Figure 1. Frames from a very low activity and a very high activity clip.

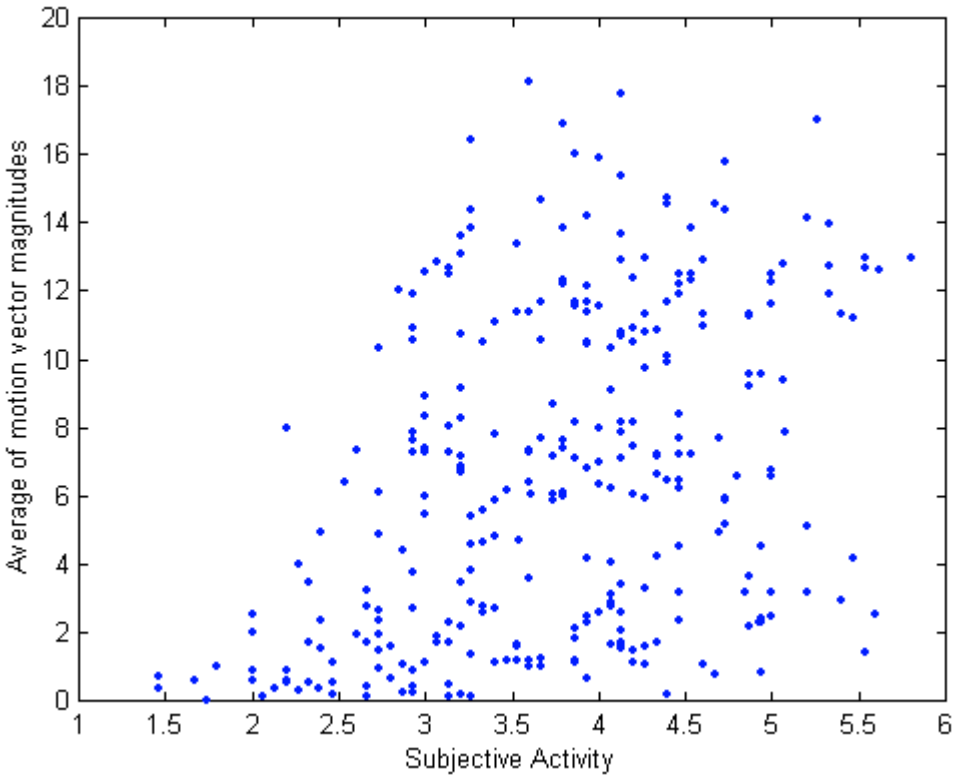


Figure 2. The scatter plot of the descriptor *avg* (average of motion vector magnitudes) vs. the average of subjects' evaluations for 294 clips. The scatter is not well structured enough for curve fitting.

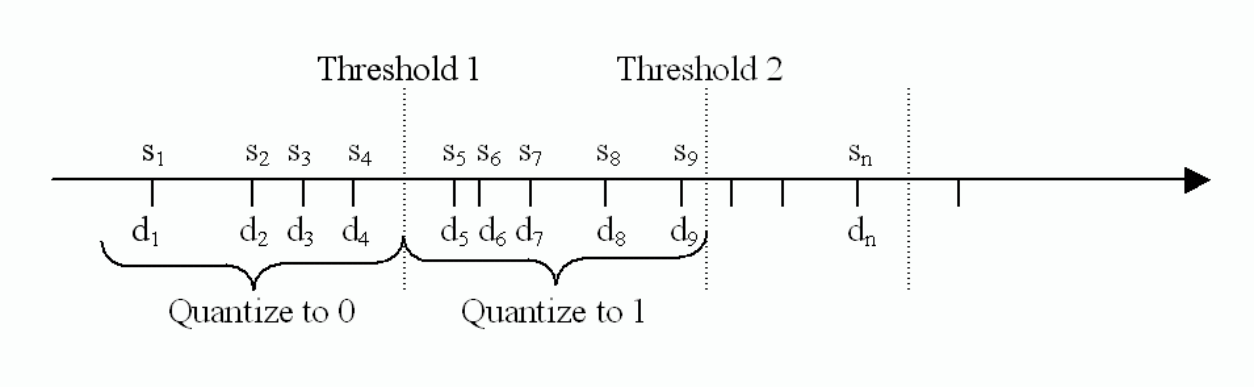


Figure 3: The quantization of a computed measures can be formulated as dividing the sorted set of N measure values into 6 by placing 5 dividers into N+1 available slots. The optimization criterion is the sum of absolute differences between the quantized descriptor values and the subjective ground truth.

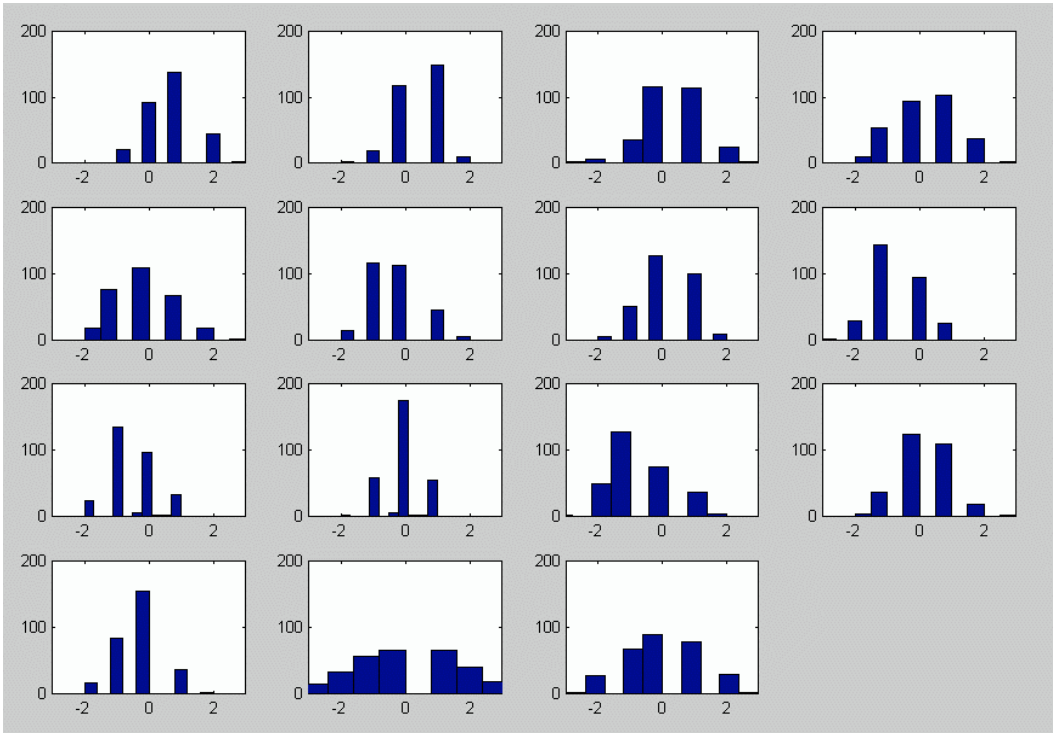
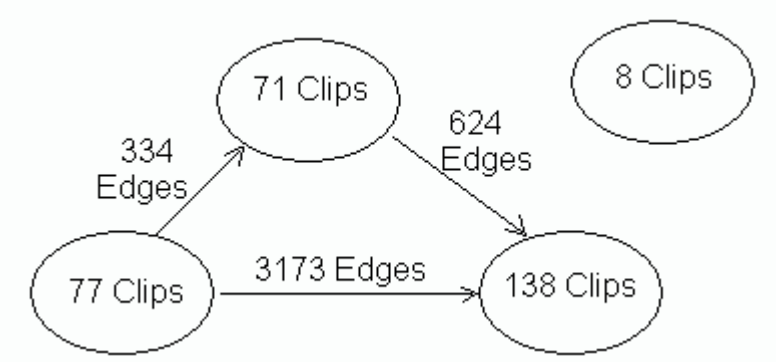
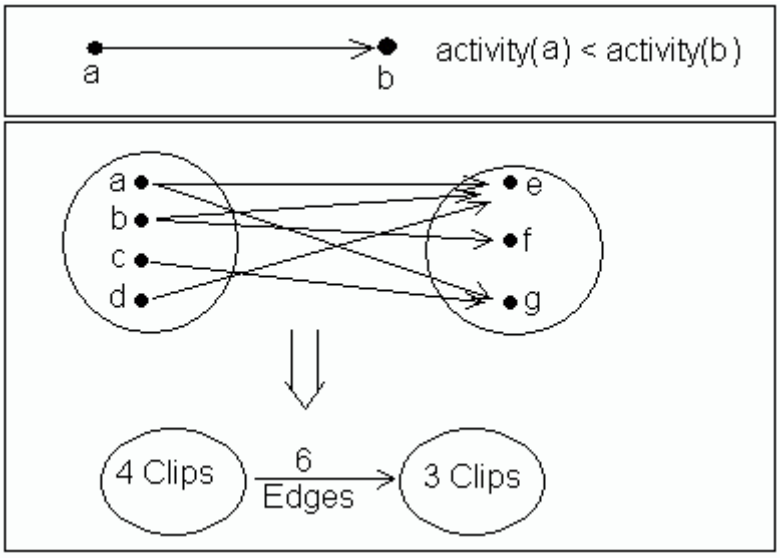


Figure 4. Histogram of each subject's errors with respect to the ground truth. Most of the subjects have a bias either towards assigning higher activity levels than the others, or lower.



(a)



(b)

Figure 5. (a) The relationship graph that illustrates the topology of the relationships established between video clips in terms of their activity level. We can identify 4 sets of clips: 1) 77 clips that are lower activity than some others, and higher activity than no other clip; 2) 71 clips that are higher activity than some and lower activity than some; 3) 138 clips that are higher activity than some and lower activity than no other clip; and 4) 8 clips that have no clear relationship with any other clip. (b) Explanation of the notation used in the graph (a).

(a)

(b)

Figure 6. Frames from examples of video clips where all motion activity descriptors fail: (a) Two clips where the motion vectors are small or cover a small area in the frame, but the perceived activity level is high. (b) Two clips where the motion vectors are very large and the motion has large frame coverage but the perceived activity level is low.