Exploiting Visual Cues in Non-Scripted Lecture Videos for Multi-modal Action Recognition

Ali Shariq Imran[1], Alejandro Moreno[2], Faouzi Alaya Cheikh[1]
[1]*Dept. of Computer Science and Media Technology, Gjøvik University College, P.O.Box-191, N-2802, Gjøvik, Norway*
[2]*University of Twente, Netherlands*
ali.imran@hig.no, a.m.morenocelleri@utwente.nl

***Abstract*—The usage of non-scripted lecture videos as a part of learning material is becoming an everyday activity in most of higher education institutions due to the growing interest in flexible and blended education. Generally these videos are delivered as part of Learning Objects (LO) through various Learning Management Systems (LMS). Currently creating these video learning objects (VLO) is a cumbersome process. Because it requires thorough analyses of the lecture content for meta-data extraction and the extraction of the structural information for indexing and retrieval purposes. Current e-learning systems and libraries (such as libSCORM) lack the functionally for exploiting semantic content for automatic segmentation. Without the additional meta-data and structural information lecture videos thus do not provide the required level of interactivity required for flexible education. As a result, they fail to captivate students' attention for long time and thus their effective use remains a challenge.**

Exploiting visual actions present in non-scripted lecture videos can be useful for automatically segmenting and extracting the structure of these videos. Such visual cues help identify possible key frames, index points, key events and relevant meta-data useful for e-learning systems, video surrogates and video skims. We therefore, propose a multi-model action classification system for four predefined actions performed by instructor in lecture videos. These actions are writing, erasing, speaking and being idle. The proposed approach is based on human shape and motion analysis using motion history images (MHI) at different temporal resolutions allowing robust action classification. Additionally, it augments the visual features classification based on audio analysis which is shown to improve the overall action classification performance. The initial experimental results using recorded lecture videos gave an overall classification accuracy of 89.06%. We evaluated the performance of our approch to template matching using correlation and similitude and found nearly 30% improvement over it. These are very encouraging results that prove the validity of the approach and its potential in extracting structural information from instructional videos.

***Keywords*-visual actions, action classification, recognition, multi-modal analysis, lecture videos.**

I. Introduction

The use of lecture videos is becoming popular as a mean of teaching for flexible education. Most of these videos cover the whole duration of a lecture with considerable pauses and inactive moments. To avoid wasting students' time watching non instructional content and storage space on servers, one needs a mechanism to automatically segment and abstract the lecture videos into small segments based on the activity it represents. Segmenting the lecture videos into small chunks makes sense if we consider the average lecture to be of 45 minutes, with lots of inactive scenes i.e. where there is nothing significant happening in the video. Finding such index points will help create efficient VLOs [1] by removing portions of inactive scenes from a video with no pedagogical value. However, creating an index for a lecture video is a challenging task [2]. One has to know what kind of indexes are good to navigate through a video. Also, finding from where to extract index points can be challenging. Instructional videos are more challenging than films or TV programs due to their non-scripted and unedited nature. The strong similarity between all frames of lecture videos and the static nature of the scenes make finding the most important sections a difficult task. One way of finding key events in such videos is to exploit the visual cues present in the instructor-led lecture videos by recognizing the lecturers' action. For example, in non-scripted videos, the actions can be useful cues to automatically segment and extract the structure of the videos. This could in turn be the first step towards a system that could be used in many different applications such as: Multimedia Learning Objects (MLOs) creation [3], video indexing, summarization or the design of smart environments [4]. In this paper, we therefore, address the classification of human action with respect to lecture videos.

Exploiting visual cues through action recognition is a challenging problem. The task is very challenging due to the vast number of actions that a lecturer can execute in different environments, conditions, etc. Classification of human motion itself has been a topic of profound interest to researchers in many different fields of computer science. This general interest in human motion stems from the fact that its potential applications are very diverse, ranging from surveillance [5] to human activity recognition [6], [7] to human computer interaction [8]. The study of human motion is not a trivial task as it is dependent on many factors, among which we can list the action that is being performed and the style (i.e. motion signature) specific to each individual [9], movement speed, posture and relative ordering of the actions being performed [10] etc. Since actions can vary greatly, it is evident that some actions may need more detailed information than others to be classified accurately. Analyzing actions using different time scales which inherently have different levels of detail would prove beneficial.

978-0-7695-4911-8/12 $26.00 © 2012 IEEE
DOI 10.1109/SITIS.2012.12

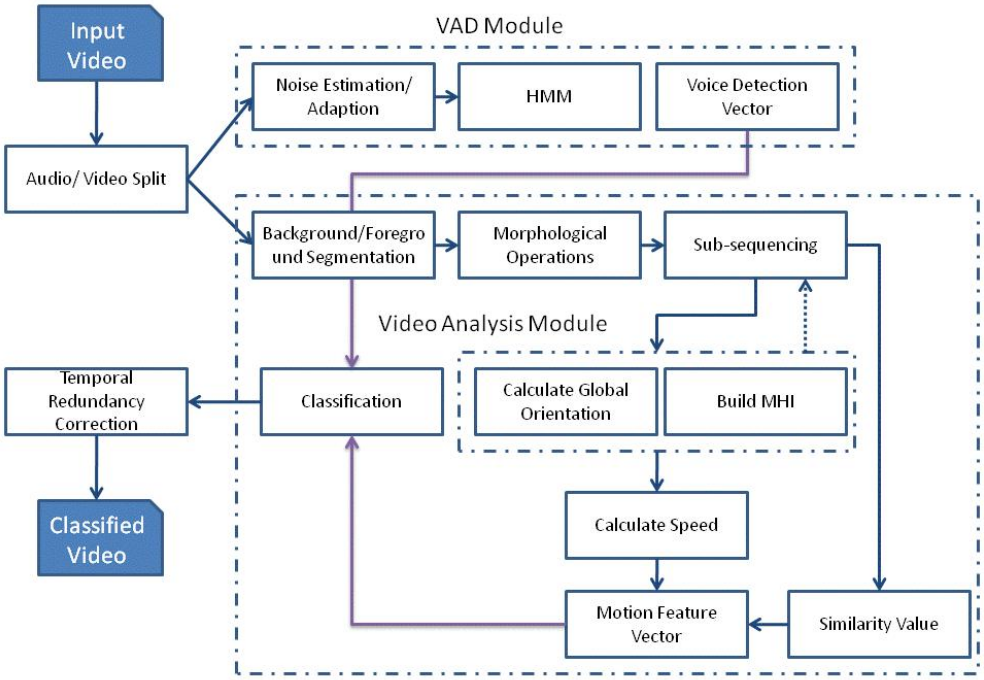

Figure 1. Overview of the proposed system.

The rest of the paper is organized as follows: in the following section we present a review of the literature related to human motion and activity recognition. Section III describes the processing steps of our proposed approach. In section IV, we present the results. In section V, we compare our approach to template-based approach. Finally, section VI concludes our paper.

II. Related Work

Babu and Ramakrishnan propose recognizing human actions based on motion information in compressed videos [11]. They do so by constructing motion history images out of motion compensation vectors present in compressed videos. Robertson and Reid present a general human activity recognition method using both position and motion information [12]. They use Hidden Markov Models to classify actions using training actions previously stored in a database. Qian et al. propose a framework for human activity recognition that uses multi-class SVM classifiers [13]. Where they use motion energy images contour coding to represent human motion. Perera et al. present a method to extract key poses from dance sequences using motion energy flow in [14]. They also propose a method for the reconstruction of low dimensional motion from high dimensional motion based on their extracted key poses. Rivera-Bautista et al. [15] show how motion information can be used as a measure of an individual's intention or attitude and programmed an autonomous robot that can decide whether to initiate interaction with individual. Ji and Liu put together a survey on view invariant human motion analysis in [16]. They state the importance of human motion research due to the many promising applications and provide a coherent overview of state of the art techniques in human detection, representation, estimation and behavior understanding.

Niebles et al. present a framework for studying motion by using the temporal structure of human actions [17]. They represent activities as groups of motion segments, where the appearance of each segment is modeled. They consider the global temporal level by modeling distinctive statistics, and the shorter temporal ranges where the patterns of the motion segments are modeled. Du et al. propose an activity recognition system using multi-scale motion detail analysis in [18]. In their approach, they refer to motion details as different components to analyze; namely motion details of trajectories, motion details of silhouettes and motion details of body parts. They analyze different scales of details or levels of abstraction, but always in the same temporal scale. Davis uses hierarchical Motion History Images (MHIs) [19] for recognizing human motion [20]. He builds MHIs for different spatial resolutions of images by using image pyramids. His proposal shares some common ground to our approach in the sense that he also analyzes motion at different resolutions, but only spatial and not temporal ones.

III. Proposed System

In our proposed approach, we first analyze the human motion by building MHIs of the teacher silhouette. We calculate global orientation and speed from these MHIs to identify writing and erasing activities. Next, we calculate the similitude value by analyzing the shape of the silhou-

ette to identify idle state from writing and erasing. We used random forest classifier to classify these states. We then apply recursive median on the classification results to rectify any classification errors. The voice activity detection (VAD) module of our system first separates audio from the video and then analyses the audio for speech detection. It identifies speaking and idle state. At the end, we get each one second sequence of a video labeled as either writing, erasing, speaking or idle. As one can speak while writing or erasing so we use two additional labels as writing + speaking, and erasing + speaking. The system consists of 2 main modules with 5 components. Figure 1 shows the overview of the proposed system model. The system components are described below.

A. Voice Activity Detection

The purpose of VAD is to recognize the presence or absence of voice in a video. As a pre-processing step to VAD, we split audio from the video using FFmpeg [21]. The audio is then processed for voice detection. The VAD algorithm used in our application is based on a statistical model-based voice activity detection [22], with additional feedback process to improve estimation done at the starting stage. This improves the accuracy significantly in environment with non-stationary noise. The algorithm returns audio frames labeled as either 0 or 1, where 1 represents presence of voice and 0 an idle state.

B. Foreground Extraction

The goal of foreground extraction is to extract moving persons from the background for motion analysis. In the first step of foreground extraction, a foreground model is created to discriminate between the foreground and background pixels. We have used the foreground object detection model proposed by Liyuan et al [23]. It detects and segments foreground objects from a video which contains both stationary and moving background objects. The stationary object is described by the color feature and the moving object is represented by the color co-occurence feature [23]. For our application we treat blackboard text and illumination changes as part of the moving background. Foreground object is extracted by fusing the classification results from the stationary and moving pixels. This model is based on Bayes theorem to classify foreground objects from background objects. Any pixel that does not fit this model is then deemed to be background.

By using the Bayes theorem, a posterior probability of v_t (i.e. feature vector extracted from an image sequence $Is(x, y)$ at time t), can be expressed as:

$$P(C|v_t, Is) = \frac{P(v_t|C, Is)P(C|Is)}{P(v_t|Is)},\ C = b\ or\ f \quad (1)$$

Therefore, by using Bayes decision rule, a pixel can be classified as foreground or background if:

$$P(v_t|Is) = P(v_t|f, Is).P_f + P(v_t|b, Is).P_b, \quad (2)$$

Figure 2. (a) Original video frame (b) Segmented foreground object (c) Frame after morphological operations and smoothing.

where $P_f = P(f|Is)$, and $P_b = P(b|Is)$.

Thus by learning a prior probability P_f, the probability P_b and conditional probability $P(v_t|f, Is)$ in advance, we can classify a feature vector v_t as either associated with background or foreground.

Quite often the extracted foreground frame contains some text noise. This is due to the non-static nature of the text during writing phase. As a result, the foreground frame contains some portions of the background text. To overcome this problem, we apply morphological opening operation on the foreground object F with structuring element S as:

$$F \circ S = (F\theta S) \oplus S, \quad (3)$$

where S is a disk shaped structuring element of radius 3. We then apply Gaussian smoothing filter to smooth out the ragged edges. The results are shown in Figure 2.

C. Motion Feature Extraction

The main idea behind the use of motion is that, even though writing and erasing are different actions, the poses to execute them are the same or very similar. Knowing this, training to differentiate between them using their postures would not work. Nonetheless, when analyzing the movement of these two actions, we see that there is an obvious difference: erasing is a back and forth continuous movement; on the other hand, writing is composed of linear movements in a certain direction. Therefore, by representing this difference of movement we can build effective motion feature vectors to characterize the different actions.

At first, we split the video into 1 second sequences that are further divided into sub-sequence of 1/2 and 1/4 of the duration as shown in Figure 3. As we build the MHI for a given sub-sequence, we also calculate the direction of the movement in the MHI. To do this, we first calculate the motion gradient at each 3x3 neighborhood of the MHI. We use the partial derivatives in the x-axis $(dMHI/dx)$ and y-axis $(dMHI/dy)$ as follows:

$$Orientation(x, y) = \arctan\left(\frac{\frac{\mathrm{d}MHI}{\mathrm{d}y}}{\frac{\mathrm{d}MHI}{\mathrm{d}x}}\right). \quad (4)$$

To calculate the orientation accurately, we need to define two thresholds: T_1 and T_2 where $T_1 > T_2$. These thresholds control the time range (i.e. frame number) in which movement in a given area is taken into account to calculate the orientation (e.g. old movement should not be taken into account). For a given neighborhood, if $M(x, y)$

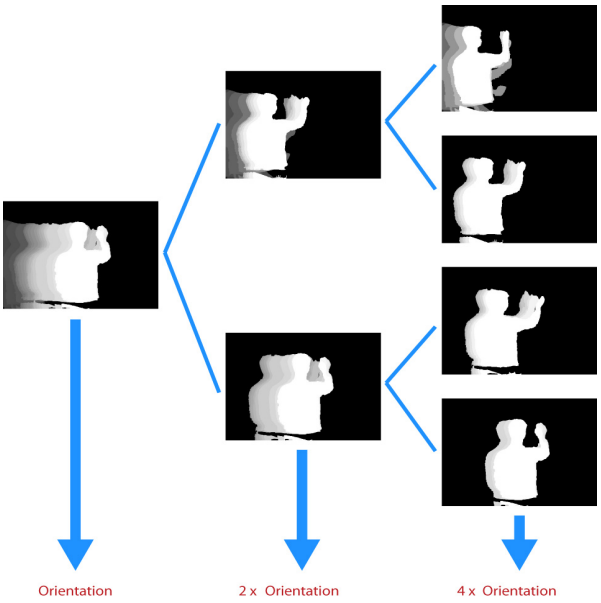

Figure 3. Level two sub-sequence applied on a MHI.

and $m(x, y)$ are the maximum and minimum MHI values respectively (timestamps), at position (x, y) the motion orientation value is only valid if,

$$T_1 \leq M(x, y) - m(x, y) \leq T_2. \tag{5}$$

Using Equation 4 and 5, we calculate the average motion direction in a selected region by building a weighted orientation histogram where recent movement has a higher weight than motion that occurred in the past. We then calculate the global orientation of the whole body movement by taking the median of the angles instead of averaging, to reduce the effect of outliers. The global orientation is calculated at each frame at the same time as we update the MHI. Since the global orientation can take any value between $0 - 360$ degrees, we decided to classify them as one of 4 general directions as shown in Figure 4, since we are interested in coarse movements for our particular classification task.

We built a 4 bin orientation histogram and filled it with the new converted values of the global orientation. Lastly, the values were normalized by dividing them by the number of total samples used in the histogram. The result is an array of size 4 which stores the normalized global orientation values. For our particular application, we only needed the horizontal direction bins since for either writing or erasing we only consider horizontal movement. The feature vector is then composed of these 2 values for all the sub-sequences created during time scaling. Thus, for level 1 sub-sequences our feature vector is composed of 4 features, and for level 2 sub-sequences it is composed of 8 features.

We calculate also relative value of the speed of object movement in each sequence by analyzing each of the frames sequentially. If N is the number of frames present in a given sequence S with height H and width W and I_n

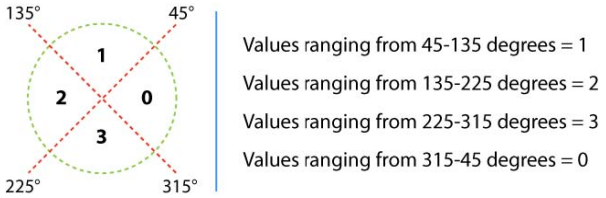

Figure 4. Mapping of global orientation values into 4 coarse movement directions.

is the image of S at frame n, the formula for the movement in S is:

$$Movement(S) = \frac{\sum_{n=2}^{N} P_D(I_n, I_{n-1})}{H * W * (N - 1)}, \tag{6}$$

where P_D is the sum of pixel values in I_n. This gives us a normalized value in $[0, 1]$, where 0 means that there has been no change in the sequence, and 1 means all the pixels have changed between a frame and the next. This value is used as a threshold in the classification step to decide whether there has been enough movement to try to classify the sub-sequence as one of the defined actions. If the speed is below the defined threshold (11% of the total number of pixels in the sub-sequence) then no classification is done on this sequence and it is labeled as an 'Idle' sub-sequence.

D. Shape Feature Extraction

The posture or shape of the silhouette has to be taken into account to distinguish writing and erasing from any other action that the teacher might be performing while giving a lecture. As the action of writing or erasing generally has a very big spatial activity area, we therefore separate the posture into three classes: high, mid and low writing/erasing action. By doing so, we can accurately differentiate idle state from writing and erasing. We then use the Hu Moments Shape Descriptor to evaluate the similitude between the shape of the silhouette and our writing / erasing silhouettes. For each frame of the sequence, we obtain the Hu Moment of the silhouette and store its similitude to the training set. Once the sequence has been processed, the similarity value is used as a threshold to verify if a particular sequence should be classified or not in the classification component. Posture discrimination between the three classes and the general idle states can be seen in Table I.

Table I
MAHALANOBIS DISTANCES FOR THE SHAPE ANALYSIS.

Writing/Erasing Pose	Low	Mid	Up
Low	**2**	13	41.5
Mid	10	**1.5**	10.5
Up	381.5	139.25	**4.75**
Idle	353	285.167	700.5

It is clear that for our three classes, we can define a valid threshold to differentiate them from the idle states by always selecting the smallest distance to one of the three classes in a given frame.

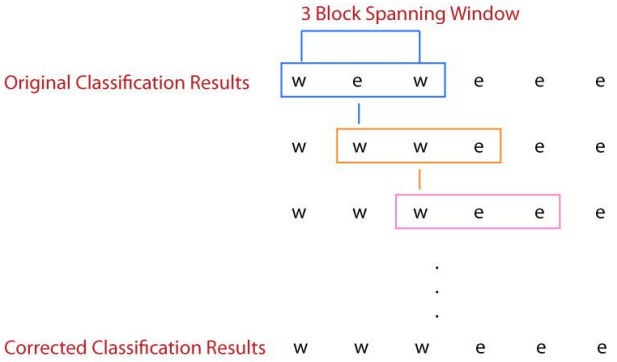

Figure 5. Recursive median on classification results.

E. Classification Based on Extracted Features

This component classifies a sub-sequence as either writing or erasing action based on extracted features. The initial feature vector consists of 8 orientation features and 4 speed of object movement features extracted from object motion at level-2 sub-sequence, along with a shape and audio feature. Later, we discard the speed and shape feature due to their low discriminative power and use them as thresholds to verify if a particular sub-sequence meets the requirement to be classified. Otherwise, the sub-sequence is classified as idle. The classifier that is used in the classification process is the Random Forest Classifier [24] as it is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier [25].

To further improve the accuracy of the classification, we take advantage of temporal redundancy in videos. We apply recursive median algorithm to correct the classification errors. The correction is applied using a 3-block spanning window that will move over the whole classification results vector. The algorithm uses the first and last classification results within the spanning window to establish whether the current result is correct or not. If both the previous and next element is of the same label, then the current element is corrected and converted to this label. If the surrounding elements are not of the same label, then there is no correction performed and the current element stays as is. Figure 5 shows an example of recursive median technique used in our approach.

IV. Experimental Results

The output of the integrated system is a video where every second has been classified into one of the predefined actions. Since the actions of talking / silence might occur simultaneously with the writing / erasing ones, we decided that talking would have precedence over the latter. The idle state is selected only when there is no recognized movement from the teacher and there is no voice detected.

The training dataset was built manually by selecting 1 second action sequences from recorded lecture videos. We collected 72 writing and 47 erasing sequences to train our random forest classifier. To test the discrimination power of our motion feature vectors (MFVs), we used five short action videos which ranged from 45 to 60 seconds. As stated before, the background segmented videos were separated into 1 second sub-sequences and processed sequentially. Three of these videos were solely writing actions and the other two were only erasing. We split the sequence up to level-2 and fuse the feature vectors obtained at level-1 (i.e. 1/2 of a sequence) and at level-2 (i.e. 1/4 of a sequence). Then we applied temporal redundancy correction (TRC) on the motion feature vector to correct isolated classification errors.

Table II
Classification accuracy using TRC on the MFVs.

	No correction	Running Median	Recursive Median
Erasing 1	60%	67%	**77%**
Erasing 2	69%	78%	**88%**
Writing 1	65%	68%	**73%**
Writing 2	63%	74%	**79%**
Writing 3	59%	**69%**	55%

The results can be seen in Table II. We then tested our VAD algorithm by using 15 audio sequences ranging from 5 - 15 sec for silence, speech, and conversation audio files. We obtained 100% accuracy for silence, 97.7% accuracy for speech and 93.6% accuracy for conversation. The overall accuracy of the algorithm was 97.1%.

To test the overall classification performance, we recorded 5 videos with 3 different teachers with no imposed limitations. The videos varied from 30 seconds to two minutes in length. To find out the ground truth, each one second sequence of every video was manually labeled. After this batch of videos had been classified, some system limitations were identified. Problems were seen when the subject arm movement was undermined by body movement. Most of the time, the subjects moved their body significantly while their arm movement was small and concise. This deceived the motion analysis component due to the greater area of the body in the silhouettes. Taking this into account, 5 more videos were recorded with 3 users where only one of them also took part in the first batch of recorded videos. In this new batch, big body movements were omitted while performing writing or erasing action. The obtained results for batch 1 and batch 2 videos can be seen in Table III.

Table III
Classification results on test videos.

Test Videos	Batch-1 video	Batch-2 video
Video 1	85.00%	93.48%
Video 2	40.63%	89.36%
Video 3	60.00%	82.83%
Video 4	50.00%	87.18%
Video 5	70.67%	92.47%
Overall	61.26%	89.06%

It can be seen that the difference when taking into consideration body movement is significant. The overall classification accuracy of the second batch of videos was 89.06%, a 27.81% improvement in comparison to the

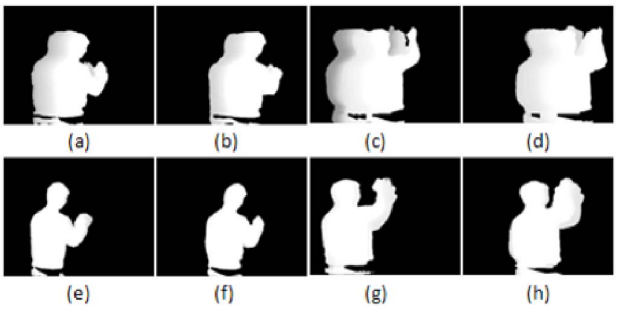

Figure 6. (a) (b) MHI of writing posture (c) (d) MHI of erasing posture (e) (f) writing templates (g) (h) erasing templates.

previous batch. This is the result of limiting big body movements while already performing a recognizeable action. To avoid imposing any limitations on subjects, we further processed the Batch-1 videos by segmenting the body into different parts and achieved 90% classification accuracy. However, these additional segmentation steps were computationally expensive.

V. Performance Evaluation

It was difficult to compare our approach to state of the art technique since our proposed approach uses multi-modal features. Nevertheless, we compared our approach to commonly used template matching techniques. For this, we created 2 types of templates as shown in Figure 6.

The templates were created by centering the MHIs of writing and erasing postures. Figure 5(a) and 5(b) are the MHIs of writing posture while Figure 5(c) and 5(d) are MHIs of erasing posture. Figure 5(e) and 5(f) are the templates created from the MHIs of writing sequence and Figure 5(g) and 5(h) are from MHIs of erasing sequence. These templates were then matched with the Batch-2 video frames to compute the classification accuracy using correlation. Let Fp and Tp denote foreground pixel and template pixel respectively. The correlation can then be defined as:

$$C_j = \frac{1}{N}\sum_{i=1}^{N} Fp_i.T_jp_i, \quad N = W * H, \tag{7}$$

where j is the total number of templates. W and H is the width and height of the video frame.

A frame is classified as either writing or erasing if the correlation is greater than certain threshold. If a threshold is below a certain level (70 for the given example) then the frame is labeled as an idle frame, according to given equation:

$$L_f = \left\{ \frac{writing/erasing}{idle}, \frac{C_j < T}{otherwise} \right\}, \tag{8}$$

where Lf is the labeled frame and T is the threshold.

We also computed the similitude value to compare the templates with the extracted video frame. This gives us slightly better results than correlation based matching criteria. Furthermore, we trained the classifier on our extracted feature using support vector machine (SVM) and tested them on Batch-2 test videos. The obtained results are very similar to random forest classifier as shown in Table IV, which proves the validity of our proposed approach.

VI. Conclusion

In this paper, we proposed a multi-modal action classification system for lecture videos. We classified four actions as writing, erasing, speaking and idle. We used these actions in media analysis and processing unit (MAPU) to identify key events for developing effective video learning objects (VLO). The MAPU is a part of our Multimedia Learning Objects (MLO) framework for E-Learning [4]. We used a multi-modal approach which is not commonly pursued in the action recognition domain. We included a VAD component to make use of audio cues that were proven to provide good classification performance improvement. We used Hu Moments as shape descriptors and the Mahalanobis distance to measure the distance to a predefined set of poses dataset. We independently analyze shape, motion and audio of the video and use the three components to obtain a more robust classification. With the presented approach, we were able to classify actions into four predefined action classes with an overall classification accuracy of 89.06%. A 25%-30% improvement over classical template matching. The proposed solution will be tested in near future with different action categories and other multi-modal single step classifier for identification of possible key events.

References

[1] X. Mu, "Decoupling the information applicaiton from the information creation: Video as learning objects in three-tier architecture," *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 1, pp. 109–125, 2005.

[2] T. Liu and J. Kender, "Lecture videos for e-learning: current research and challenges," in *IEEE Sixth International Symposium on Multimedia Software Engineering*, dec. 2004, pp. 574 – 578.

[3] A. S. Imran, "Interactive media learning object in distance and blended education," in *Proceedings of the 17th ACM international conference on Multimedia*, ser. MM '09. Beijing, China: ACM, 2009, pp. 1139–1140.

[4] A. S. Imran and F. Alaya Cheikh, "Multimedia learning objects framework for e-learning," in *The International Conference on E-Learning and E-Technologies in Education, Lodz, Poland*. ICEEE, 2012, pp. 105–109.

[5] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334 –352, aug. 2004.

[6] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224 – 241, 2011.

Table IV
COMPARISON BETWEEN TEMPLATE MATCHING AND PROPOSED APPROACH (OBTAINED CLASSIFICATION ACCURACY IN % ON 5 TEST VIDEOS FROM BATCH-2).

Test Videos	Template Matching using	Proposed Approach using
Batch-2	Correlation — Similitude	SVM — Random Forest
Video 1	67% — 69%	91% — 93%
Video 2	43% — 55%	90% — 89%
Video 3	58% — 51%	85% — 82%
Video 4	50% — 57%	83% — 87%
Video 5	63% — 64%	88% — 92%
Overall	56% — 59%	87% — 89%

[7] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976 – 990, 2010.

[8] A. Nijholt, M. Pasch, B. van Dijk, D. Reidsma, and D. Heylen, "Observations on experience and flow in movement-based interaction," in *Whole Body Interaction*, ser. Human-Computer Interaction Series, D. England, Ed. London: Springer-Verlag, May 2011, pp. 101–119.

[9] M. A. O. Vasilescu, "Human motion signatures: analysis, synthesis, recognition," in *Proceedings of 16th International Conference on Pattern Recognition*, vol. 3, 2002, pp. 456 – 460 vol.3.

[10] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Underst.*, vol. 104, no. 2, pp. 90–126, Nov. 2006.

[11] R. V. Babu and K. R. Ramakrishnan, "Recognition of human actions using motion history information extracted from the compressed video," *Image and Vision Computing*, vol. 22, no. 8, pp. 597–607+, 2004.

[12] N. Robertson and I. Reid, "A general method for human activity recognition in video," *Computer Vision and Image Understanding*, vol. 104, pp. 232–248, 2006.

[13] H. Qian, Y. Mao, W. Xiang, and Z. Wang, "Recognition of human activities using svm multi-class classifier," *Pattern Recognition Letters*, vol. 31, no. 2, pp. 100 – 111, 2010.

[14] M. Perera, S. Kudoh, and K. Ikeuchi, "Keypose and style analysis based on low-dimensional representation," *IPSJ SIG Notes. CVIM*, vol. 2009, no. 2, pp. 1–16, 2009-06-02.

[15] J. A. Rivera-Bautista, A. C. Ramirez-Hernandez, V. A. Garcia-Vega, and A. Marin-Hernandez, "Modular control for human motion analysis and classification in human-robot interaction," in *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2010*, march 2010, pp. 169 –170.

[16] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 1, pp. 13 –24, jan. 2010.

[17] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proceedings of the 11th European conference on Computer vision: Part II*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 392–405.

[18] Y. Du, F. Chen, W. Xu, and W. Zhang, "Activity recognition through multi-scale motion detail analysis," *Neurocomputing*, vol. 71, no. 1618, pp. 3561 – 3574, 2008, advances in Neural Information Processing (ICONIP 2006) / Brazilian Symposium on Neural Networks (SBRN 2006).

[19] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257 –267, mar 2001.

[20] J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp. 39 –46.

[21] FFmpeg, "FFmpeg," http://www.ffmpeg.org, accessed 30-Aug-2012. [Online]. Available: http://www.ffmpeg.org

[22] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," vol. 6, no. 1, pp. 1–3, 1999.

[23] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the eleventh ACM international conference on Multimedia*, ser. MULTIMEDIA '03. New York, NY, USA: ACM, 2003, pp. 2–10.

[24] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[25] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 96–103.