IDENTIFYING SALIENT POSES IN LECTURE VIDEOS

John R. Zhang, John R. Kender

Columbia University
Dept of Computer Science
New York, USA

ABSTRACT

The communicative importance of gestures in teaching environments have been widely studied. Two classes of gestures—*point* and *spread* gestures—have been identified to indicate pedagogical importance in teaching discourse [1]. In this work, we propose a system for the identification of the poses of point and spread gestures as a preliminary step toward their identification in low-quality unstructured videos. We use a joint-angle descriptor derived from an automatic pose estimation framework to train an SVM in order to classify extracted video frames of an instructor giving a lecture. Ground-truth is collected in the form of 2500 manually annotated frames covering approximately 20 minutes of a video lecture. Cross validation on the ground-truth data showed initial classifier F-scores of 0.54 and 0.39 for point and spread poses.

***Index Terms*—** Image classification, lecture videos, gestures, poses.

1. INTRODUCTION

The importance of gestures in lecture settings has been widely studied. Existing work by Roth et al. [2, 3] have shown that certain gestures by instructors can serve as semantic cues and have a direct effect on students' learning. In this paper, we propose a system for the identification of salient poses in lecture videos as a preliminary step towards identifying salient gestures, which may ultimately be used as semantic cues in non-linear semantic video browsers such as Vast MM [4].

We focus on the domain of lecture videos. The increasing prevalence of online video has greatly increased the practice by institutions to digitally record lectures and make them publicly available on the Internet[1]. These recorded lectures provide an invaluable source of knowledge for both students of these institutions as well as the general public. However, they also raise the challenge of information overload. Non-linear semantic video browsers pose a possible solution to this problem by using multimedia cues from video, audio and text to guide viewers to segments of the video which may be of particular interest. In particular, gestures of enphasis (arm pointing) and gestures that iconify the difficulty of material (arms spreading) appear especially valuable as video indices.

The paper is divided into the following sections. In Section 2, we briefly review existing research on the relevance of human gesture in the teching context as well as work in the computer vision field for the estimation and identification of poses. In Section 3, we describe our system for identifying frames of video containing an instructor in a semantically relevant pose (as determined by the literature). In Section 4 we test our system on a university computer science lecture which has been manually annotated for both the target poses as well as analyzed for semantic significance. Finally, we conclude in Section 5 with highlights of our contributions and a discussion of future work.

2. RELATED WORK

A number of existing works in education and psychology have examined the importance of the gestures in communication.

In [2], Roth et al. examine the importance of hand and arm gestures relative to body position and motion. The authors find that gestures can sometimes convey information that is not conveyed in speech alone, as well as the discovery that some children could express understanding of taught material through gesture if not speech. This work hints at the feasiblity of the usage of gestures as semantic cues. It also argues for the prominent role of hands, arms and body in relevant gestures. We apply this to our work by focusing on these body parts as an initial step.

Much work has also been done on the taxonomy and representation of gestures. The work of Kendon argues for the use of a *tri-phasic* model for gestures [5]. In this model, gestures can be separated into three phases: a position of rest (*preparation*), a peak structure (*stroke*) and a return to a position of rest (*retraction*). We employ this model to represent gestures in our application. Futhermore, as we do not yet employ temporal features in our system, we identify each gesture using just their stroke pose (selected as a single frame of video in a subsequence denoted as belonging to a gesture).

[1] e.g. http://www.cvn.columbia.edu or iTunes U

Zhang et al. analyzed gestures in lecture videos and proposed a taxonomy dividing gestures semantically relevant to teaching into nine classes [1]. Two of these classes, *point* and *spread*, were found to occur especially frequently. Point gestures were the most frequently occuring, as instructors often pointed to refer to specific topics on a blackboard or slide (see Figure 1a,b). When an instructor points, it is usually a cue for students to pay attention to a particular topic. Spread gestures usually involved both hands and arms extended in front of the body (see Figure 1c,d). Gestures of this type were found to serve as pedagogic commentary, independent of lecture content, with amount and duration of spread indicating the difficulty of the material. Together, point and spread gestures accounted for over 50% of the gestures in the videos examined. Due to their prominence, we focus on identifying the poses associated with these gestures here.

For the task of automatic pose estimation from an image, we review the work of Ferrari et al. [6]. In this work, the pose of a person in an image (modeled using six body parts: head, torso, upper and lower left and right arms; the four leg parts are omitted) is extracted. This is achieved through a multi-phase approach whereby the position, orientation and scale of each part is estimated. An initial weak model is applied to roughly locate the torso and head, thereby reducing the search space for a stronger model. Full body detection, segmentation and then iterative parsing is used to produce the final pose. The method is robust, fully automatic and self-initializing. We use this pose estimation method in our work.

Ferrari et al. also examined the problem of pose search [7], and proposed a system for the retrieval of certain poses. The authors proposed 3 descriptors which can be used in 2 approaches: input can be taken in the form of an example pose and then similar poses (according to a provided similarity metric) are returned, or alternatively, sample poses can be used to train a classifier. Here, we develop an alternate descriptor based on joint angles which is simple to compute and is able to capture the target poses in principle (i.e. they cover the used body parts).

3. METHOD

We propose a preliminary system toward the goal of identifying gestures salient to teaching in lecture videos. The system presented here will allow for the identification of the stroke poses of the gestures discussed in Section 2.

Given a lecture video as input, the frames are extracted at a constant rate. We can make some domain-specific assumptions which simplify the problem. We assume that the video focuses on a single person (i.e. the instructor) who tends to be near the center of the video. If multiple persons are detected, whether it is because a student was caught in the frame or because of a false detection, then we can assume that the one with a midpoint closest to the center of the image is that of the instructor. A second assumption is that the person of interest is always standing upright, that is, the torso is vertical (with a small range of flexibility). Anything exceeding this range can be assumed to be a poor estimation.

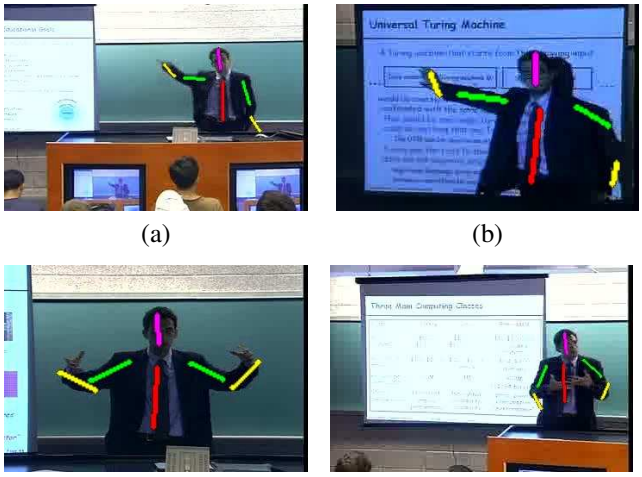


Fig. 1. Examples of point poses (a, b) and spread poses (c, d) with automatically estimated poses overlayed.

As an initial step, we apply a human detector on all frames and disregard those with negative results.

Next, for the actual task of pose identification, we aim to classify each input image as belonging a point pose, a spread pose, or neither. We approach the problem by building binary classifiers for each class. We elaborate on the construction of these classifiers as follows.

3.1. Pose Estimation

The pose of the instructor in focus is the basis for our descriptors. As discussed in [2, 1], the arms and torso play a significant role in gestures in teaching. Therefore, we focus on five body parts: the torso, lower and upper right arm, lower and upper left arm.

To extract the approximate positions and orientations of these body parts from a given image, we use the pose estimation framework introduced by Ferrari et al. [6]. This framework provides the endpoints of each body part in question as output (among other data). Figure 1 shows the output of the pose estimator overlayed on sample input images. Examples of imperfect pose estimations are shown in Figure 2.

The extracted pose is then represented as a set of 2D vectors pointing outwards from the neck (see Figure 3). To get the appropriate directions, we begin by assuming that the torso is more or less vertical, and then associate the related joints by finding the closest points. Given this vector representation of the five body parts, we can compute a descriptor comprised of joint angles.

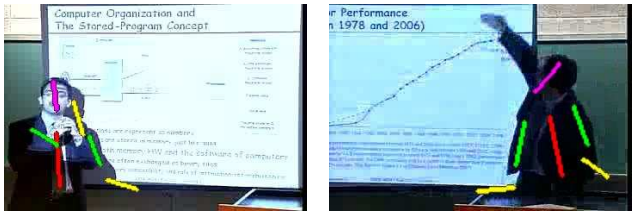


Fig. 2. Examples of images where automatic pose estimation was imperfect.

3.2. Pose Descriptors

For each pose detected in an image, we can compute a 4-dimensional descriptor $D = [\alpha_0, \alpha_1, \alpha_2, \alpha_3]^T$. The values α_i represent the angles between the torso and left upper arm, torso and right upper arm, left upper arm and left lower arm, and right upper arm and right lower arm for $i = 0, \ldots, 3$ respectively (see Figure 3).

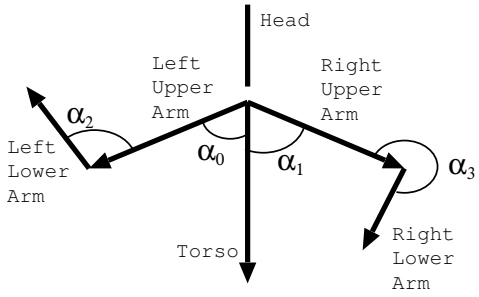


Fig. 3. Model and descriptor values of estimated poses. Each body part is represented as a vector (depicted as arrows; the head is ignored).

Given vectors A, B representing a body part and the body part that is connected to it, we can compute the angles α_i as follows.

$$\alpha_i = a_i \cos^{-1}\left(b_i \frac{A \cdot B}{|A|\,|B|}\right)$$

where

$$a_i = \begin{cases} -1 & \text{if } i \in \{0,2\} \text{ and } B \text{ is above } A \\ -1 & \text{if } i \in \{1,3\} \text{ and } B \text{ is below } A \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

$$b_i = \begin{cases} 1 & \text{if } i \in \{0,1\} \\ -1 & \text{otherwise} \end{cases} \quad (2)$$

Note that for Equation 1, the determination of whether or not vector B is above or below A can be done in a variety of ways and is omitted here for conciseness.

3.3. Classifier

We use LibSVM[2] to train a classifier. We heuristically determined that the RBF kernel performed best for both classes. Parameters were found using a simple grid search.

4. EXPERIMENTS

We evaluate our pose classification system by attempting to classify point and spread poses on manually labeled ground-truth. Approximately 20 minutes of a computer science video lecture was manually annotated. The video was sampled at 2 frames per second (as was done in [1]) resulting in 2500 frames. The video features a single instructor standing in front of both a blackboard and slide giving a lecture on computer architecture. Occasionally, student(s) sitting near the front of the classroom can be seen in the foreground. The videos were provided by the Columbia Video Network: cameras were human operated by lightly trained students in ambient light with no post processing. The video is provided at the low resolution of 352×240—audio and video quality are poor. The lighting condition varied throughout the video as the video faded in and out, or as the instructor adjusted the lights (to shift focus to and from the slides). The video does not focus solely on the instructor, as it shifts focus to a view of the slides from time to time.

4.1. Evaluation

We trained a binary classifer for each of the point and spread classes. For the automatic pose estimation, we use a pre-trained model provided by Ferrari et al. which we empirically found to produce satisfactory results.

From the 2500 frames collected for ground-truth, we identified 141 positive and 970 negative samples for the point class, and 125 positive and 986 negative samples for the spread class. The remaining frames were discarded as they did not have a visible person.

For evaluation, we perform 2-fold cross validation. The results, as measured by F-score are shown in Table 1. During training, we use all possible positive samples available in the partition but limit the number of negative samples to 7 times the number of positive samples (factor selected heuristically). This subset is randomly chosen from the entire pool of negative samples. During testing, all samples available in the testing partition are used.

Table 2 shows a confusion matrix constructed from these results.

4.2. Discussion

The performance of the classifier on identification of point poses is considerably superior than performance for identifi-

[2]http://www.csie.ntu.edu.tw/~cjlin/libsvm

Class	Precision	Recall	F-Score
Point	0.72	0.47	0.54
Spread	0.35	0.45	0.39

Table 1. Performance results of point and spread classifiers.

		Predicted		
		Neither	Spread	Point
Actual	Neither	744	77	24
	Spread	76	54	11
	Point	30	38	57

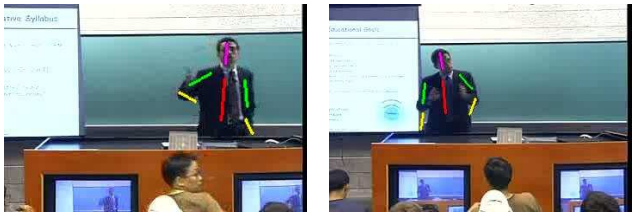
Table 2. Confusion matrix comparing classes.

cation of spread poses. This is not surprising as it is a rather distinctive pose. Misclassifications in this class tend to arise from poor pose estimations.

In this dataset, the point class is comprised entirely of pointing to the left. The reason for this is straighforward, as the instructor will tend to point to notes on the slides (he does not use the blackboard) which is always located to his right (when facing the camera). Presumably, pointing in the other direction could be trained in the same way.

Classification of spread poses was considerably more difficult. It encompasses a wider range of possible poses, some of which—e.g. standing with both arms outstretched—are difficult to distinguish in 2 dimensions. Two reasons were noticed which may account for the difficulty in training good models. One was the high number of incorrectly estimated poses, particularly of the arms, when the person is posing with arms outstretched (e.g. Figure 4d). Future work could explore the addition of hand detectors which may be used to aid pose estimation. A second reason was the granularity of the angles and the inherent noisiness of spread poses. For instance, the person with both arms straight with hands resting on the table, and the person with slightly bent elbows with hands outstretched may both produce the same pose estimation.

Figure 4 shows examples of misclassifications for both point and spread classes.



(a) False positive point. (b) False negative spread.

Fig. 4. Examples of misclassifications.

5. CONCLUSION

We have proposed and examined a preliminary system for the identification of the stroke poses of point and spread gestures in candid unstructured lecture videos, which have been identified to hold pedagogical significance in teaching discourse.

Future work would focus on extending the system to recognizing multiple frames to take advantage of temporal features and allow for gesture classification. The method also needs to be evaluated against other videos and compared against other descriptors. The system could also be extended to recognize other semantically relevant gestures as identified by [1] and others.

6. REFERENCES

[1] J. R. Zhang, K. Guo, C. Herwana, and J. R. Kender, "Annotation and taxonomy of gestures in lecture videos," in *Proc. CVPR Workshop on Human Communicative Behavior Analysis*, June 2010.

[2] W. Roth, "Gestures: Their role in teaching and learning," *Review of Educational Research*, vol. 71, no. 3, pp. 365–392, 2001.

[3] W. Roth and D. Lawless, "When up is down and down is up: Body orientation, proximity, and gestures as resources," *Language in Society*, vol. 31, no. 01, pp. 1–28, 2002.

[4] A. Haubold and J. R. Kender, "Vast mm: multimedia browser for presentation video," in *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, New York, NY, USA, 2007, pp. 41–48, ACM.

[5] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The relationship of verbal and nonverbal communication*. 1980, pp. 207–227, Mouton Publishers.

[6] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2008.

[7] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Pose search: Retrieving people using their pose," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.