LabelMovie: Semi-supervised Machine Annotation Tool with Quality Assurance and Crowd-sourcing Options for Videos

Zsolt Palotai
Colleyeder Ltd.
Budapest, Hungary H-1121
Zsolt.Palotai@colleyeder.com

Miklós Láng
Colleyeder Ltd.
Budapest, Hungary H-1121
Miklos.Lang@colleyeder.com

András Sárkány
Colleyeder Ltd.
Budapest, Hungary H-1121
Andras.Sarkany@colleyeder.com

Zoltán Tősér
Colleyeder Ltd.
Budapest, Hungary H-1121
Zoltan.Toser@colleyeder.com

Daniel Sonntag
German Res. Center for Artif. Intell.
Saarbrücken, Germany D-66123
Daniel.Sonntag@dfki.de

Takumi Toyama
German Res. Center for Artif. Intell.
Kaiserslautern, Germany D-67663
Takumi.Toyama@dfki.de

András Lőrincz
Eötvös Loránd University
Budapest, Hungary H-1117
lorincz@inf.elte.hu

***Abstract*—For multiple reasons, the automatic annotation of video recordings is challenging. The amount of database video instances to be annotated is huge, tedious manual labeling sessions are required, the multi-modal annotation needs exact information of space, time, and context, and the different labeling opportunities require special agreements between annotators, and alike. Crowd-sourcing with quality assurance by experts may come to the rescue here. We have developed a special tool: individual experts can annotate videos over the Internet, their work can be joined and filtered, the annotated material can be evaluated by machine learning methods, and automated annotation may start according to a predefined confidence level. A relatively small number of manually labeled instances may efficiently bootstrap the machine annotation procedure. We present the new mathematical concepts and algorithms for semi-supervised induction and the corresponding manual annotation tool which features special visualization methods for crowd-sourced users. A special feature is that the annotation tool is usable for users not familiar with machine learning methods; for example, we allow them to ignite and handle a complex bootstrapping process.**

I. Introduction

Annotation of videos is of great interest for content providers for monitoring, surveillance, meteorology, maritime processes and similar authoring tasks. Advanced solutions should include more precise content-based annotations by extending common data annotation tools by the (semi-) automatic annotation of video recordings. The annotations themselves serve video retrieval and browsing. Annotation can be guided by contextual and content-based information, and can rely on auditory, visual, textual, color information. Annotation can aim at more sophisticated goals, such as the annotation of player's behavior in an educational game and thus help in the personalization of educational training material. It can also assist authoring, for example. In fact, methods for media annotation to perform the whole application cycle of annotation, query and analysis are highly sought after. End user applications includes interactive narratives, 3D motion capture of humans, virtual camera movement of automatically captured 3D environments and so forth.

There is a considerable interest in smart video annotation including video browsing [1], scalable crowd-sourcing based annotation [2], and semantic and ontology based annotation [3], [4]. Elaborated annotation tools (for a precise annotation of human movement) have been developed in recent years, see, e.g., the Anvil tool [5]. Intelligent annotation using content-based multi-media information retrieval is the subject of active research and evaluation; for details on progress and results, see the National Institute of Standards and Technology sponsored conference series[1] for example.

Our annotation tool development stems from a crucial bottleneck that we have been facing during the evaluation of educational games: we expected that human annotation will be effective in searching and finding applications which need annotation-driven interaction to improve progress and keep up the level of engagement and entertainment of each individual user. In a series of experiments ran by researchers of the Eötvös Loránd University and the University of Szeged [6] we found that (i) experts' opinions about annotations can differ quite a lot (missing inter-annotator agreement), (ii) their annotations are uncertain, (iii) education experts may be unable to decide whether the number of annotated samples are sufficient or not for invoking machine annotation, (iv) experts wrongly expect the computer to 'see' and find what they see in the multimedia material (which is often not the case); and (v) , from an HCI perspective, the annotator must have an enabling tool that activates machine competencies and gives rise to a 'common goal' [7]. We decided to develop a tool that uses state-of-the-art technology for clustering and the recognition of spatio-temporal events in such a way that it would also serve

[1] http://trecvid.nist.gov/

non-experts to explore the capabilities of machine learning methods.

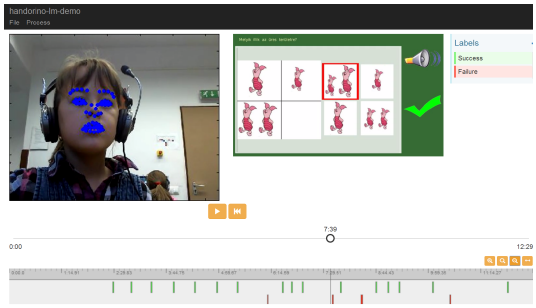
In what follows, we describe the features of the tool, three basic algorithmic solutions that we included and two demos that help describe the usage of this tool in selected new application domains.

II. LabelMovie

In this section we describe our tool and related annotation procedures for a single annotator.

A. User interface and procedure

We show the user interface in Fig. 1. Demonstration videos can be downloaded[2]. They show the two main steps of the annotation process in Fig. 1(a) and 1(b), respectively.



(a)

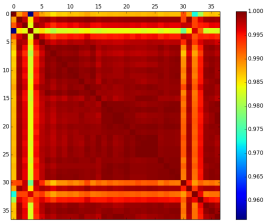
(b)

Fig. 1. Screenshots of using the annotation tool. (a) Educational game with facial markers (left) and gaming context (right). Labels are 'success' and 'failure', time resolution is low. (b) Indoor navigation with head-mounted display, camera to the environment, eye-tracking cameras with two-eye-based gaze estimation (left), 9-DOF inertia motion units on the head and on the body (right). Labels are 'standing' still and 'walking' straight, time resolution is changed to higher with the zoom tool.

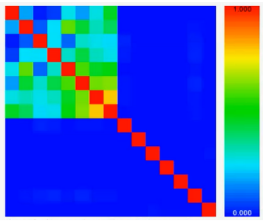
The user interface is grouped into seven interaction areas:

1) roll-off menus for file management on Google Drive, annotation processing, and crowd-sourcing functions;
2) video playback functions for watching and comparing multiple videos concurrently;
3) label list for managing the available labels, as well as their their descriptions and colors;
4) video playback buttons for controlling the (automatic) playback of the videos;

[2]http://people.inf.elte.hu/lorincz/handorino_lm_demo.mp4 and http://people.inf.elte.hu/lorincz/indoor_nav_lm_demo.mp4



(a) Gram matrix for Handorino



(b) Gram matrix for navigation

Actual / Predicted	Standing	Walking
Standing	8	0
Walking	0	7

(c) Confusion matrix for indoor navigation

Fig. 2. Gram matrix and confusion matrix. Gram matrices for (a) Handorino educational game and (b) indoor navigation examples. Samples are sufficient for automated annotations for (b) but are borderline for (a) as indicated by the confusion matrix (only shown for (b)). (c): Confusion matrix after Leave-One-Out cross validation for an indoor navigation example. Samples are sufficient since confusion matrix is diagonal.

5) a coarse-grained circle playhead for searching within the videos;
6) a zoom tool for zooming in or out given time intervals of the videos in the annotation area;
7) the annotation area with fine-grained vertical line playheads for annotating the videos (add, delete, shift annotations) and for setting the fine-grained time interval within the videos.

Automated annotations and their visualizations can be accessed in the *Process* roll-off menu. The *Gram matrix* option lets the annotator inspect the automatic annotations visually (to demonstrate the capabilities of our algorithms). If the annotator created k labels and a total of $n \geq k$ annotations with these labels then the Gram matrix will contain n columns and n rows. If the Gram matrix has k separate blocks then our algorithms can learn these annotations and can automatically annotate the remaining part of a specific video. Figure 2(a) shows a negative example; a positive example is depicted in Fig. 2(b).

Quantitative evaluation is the final step: this step can be accessed by the *confusion matrix*-option in the Process roll-off menu: a classifier is trained for the selected labels and the resulting matrix quantitatively shows the performance of the trained classifier on the test set. The classifier is perfect if the confusion matrix is diagonal (see Fig. 2(c) for this perfect case).

In summary, the annotation and evaluation procedure have the following steps (if annotations can be used for automated evaluation):

1) **Create labels:** create the labels to be used during annotation;

2) **Annotate:** annotate different events of the video with the corresponding labels;
3) **Inspect Gram matrix:** request a Gram matrix for visual inspection;
4) **Inspect confusion matrix:** request a confusion matrix to see whether the structure of the Gram matrix has blocks;
5) **Decide about automated annotation:** if the level of confusion is sufficient for your purposes then launch the automated annotation;
6) **Quality assurance:** sample the labels provided by the computer, fix them in case of an error, and/or restart automated annotation with the seeds when desired.

B. Algorithms

For the sake of completeness, we briefly describe the algorithms involved and we also provide the underlying motivations for our selection and provide the references for the interested reader. Note, that the backing algorithms are flexible and can be modified. We also present our thoughts on the possibilities of algorithmic extensions.

1) Dynamic time warping and visualization: Since annotated events can take different times, it is important to compare multiple time series of different lengths. *Dynamic time warping* (DTW) can robustly solve this problem. DTW is traditionally solved by dynamic programming.

Recently, efficient DTW procedures that utilize kernel methods appeared in the literature [8]. Kernel based classifiers are robust against invariance and distortions. LabelMovie uses the Global Alignment (GA) kernel since it shows superior performance [9].

GA kernel assumes that the minimum cost of alignments may be sensitive to peculiarities of the time series and replaces this quantity with the sum of the cost of all alignments weighted exponentially. According to the argumentation, this gives rise to a smoother measure than the minimum of these costs, and the induced Gram matrix do not tend to be diagonally dominated as long as the temporal sequences have similar lengths [8].

A Gram matrix is depicted on Fig. 2(a). Each row and column belong to an annotation. If there are n annotations of k labels, then the Gram matrix will have n rows and n columns. Columns and rows are grouped by labels. The Gram matrix shows the similarity of the annotations to each other based on the GA kernel. The annotation is good for further processing if the Gram matrix has $k \times k$ distinct blocks, or if it has k blocks in the diagonal, like in Fig. 2(b).

2) Time-series classification with Support Vector Machine: LabelMovie applies Support Vector Machine classifier [10] with GA kernel to learn the annotations.

Support Vector Machines (SVMs) are very powerful for binary and multi-class classification as well as for regression problems [11]. They are robust against outliers. For two-class separation, SVM estimates the optimal separating hyper-plane between the two classes by maximizing the margin between the hyper-plane and closest points of the classes. The closest points of the classes are called support vectors; the optimal separating hyper-plane lies at half distance between them.

LabelMovie is using '*one-against all*' classification, where decision surfaces are computed for all labels, respectively.

3) Confusion matrix: The confusion matrix shows classification mistakes on the labelled database. It works by using those annotated samples that were left out during training. Matrix elements of the confusion matrix correspond to the number of samples indexed by (i) the predicted class and (ii) the index of the correct class. In turn, for n classes, it is an $n \times n$ matrix. LabelMovie uses *Leave-One-Out* n-fold cross validation; all samples are tried as the test sample during this procedure (Fig. 2(a)-(c)).

C. Options for further extensions

LabelMovie's potential extensions include algorithmic improvements since time vector comparisons are time consuming and faster methods and/or GPU implementations could have a positive impact. Another issue concerns the classification scheme: Support Vector Machines are well documented, very efficient, and the default values for the different kernels work rather well. Novel methods (deep networks in particular) can be used since they show superior performance on many related benchmark classification task (see, e.g., [12] for the fundamentals, [13] as an example, and [14] for a review.) Last but not least, LabelMovie's design fits both crowdsourcing and quality assurance (data quality) goals of large heterogeneous (massive) datasets being promising for future multimedia indexing mechanisms for effective multimedia information extraction [15] and content-based retrieval on the large scale [16].

III. LabelMovie crowdsourcing and quality assurance

Crowdsourcing and quality assurance were key objectives in the design of LabelMovie. In our previous studies [6], [17] we found a crucial bottleneck: domain experts such as education experts, and experts of the technical (semi-) automatic annotation task (like experts of machine vision and machine learning) should join their knowledge and that this task is far from being trivial. Here, we list some of the problems we encountered followed by our solutions.

1) **Knowledge:** What is seen and recognized by the expert may not be feasible for computer based annotation. Human experts should be able to experience the capabilities and the limitations of computerized methods. Visualization of the Gram-matrix seems specifically well suited for this problem and the confusion matrix gives proper feedback about the needed sample types.
2) **Quality assurance:** As expert annotations may differ, LabelMovie works over the Internet and experts can get an overview and edit the annotations of others. Annotators may play different roles.
3) **Knowledge explanation:** Different expert opinions mean that knowledge is not firm. LabelMovie is an integration tool that (i) can combine expert knowledge

and (ii) can join information on cutting-edge transdisciplinary fields such as educational games, or robotic surgery [18]. Connections to Wikipedia, Scholarpedia, and other knowledge bases, extensions with ontology building and merging capabilities can help the explanation procedure. [18].

4) **Machine support:** LabelMovie's algorithmic components learn by examples. Sampling of machine annotations give rise to positive (i.e., true positive and true negative samples) and negative examples (i.e., false positive and false negative samples). Sampling can save time in extending previous manually-annotated sample sets.

IV. Examples

We summarize the two examples that serve to demonstrate the functioning and mode of operation of LabelMovie; the educational game and the navigation task in a building.

A. Educational game: insufficient statistics

While playing computer games, the environment of the user is well described by the state of the game. Therefore computer games fit well automated annotation since precise knowledge of the environment and the task are both known. If we can characterize the emotional, cognitive and communicative capabilities of the user, then the game can be optimized for challenges, learning trajectories, and an enhanced entertainment level. In the demonstration, automated action unit estimation [19] and emotion annotation were added to the recorded videos. Human annotation indicated specific events when human intervention is desired during learning. The sample size was low, statistics was insufficient, and the Gram matrix (Fig. 2(a)) shows the need for further human annotations.

B. Navigation and location service with smart tools

In navigation and localization tasks the environment is typically not well characterized. The more information is collected, the better the contextual description is, and the easier the machine annotation in the end. In our example, we used eye-tracking glasses with attached head mounted displays and 9 degree of freedom (9-DOF) inertial motion unit (IMU) as well as a 9-DOF IMU attached to the body to collect tracking data. The hallway had special signs that were searched for during task execution. We could thus learn the magnetic field within the hallway and analyse the quality of gaze calibration when the user stopped in front of a one of the signs. It also enabled us to estimate the head position with about 20-30 cm precision within the building. These information pieces can be used, e.g., to estimate the step sizes from gyroscope and accelerometer data. Such estimation could be exploited for trajectory estimation even in an unknown environment. Estimations can be developed step-by-step and they can be fused and updated during navigation or at calibration points, respectively. In Fig. 2(c) we show that the automated annotation of calibration points is of high quality – Gram matrix of this first step is block-diagonal and the confusion matrix showed perfect performance – and thus it can be the base of further estimations and annotations, e.g., in the eye-gaze based video material. A further evaluation of stride length may be built upon such precision of calibration points.

V. Conclusions

LabelMovie has been developed to enable domain experts without knowledge about the underlying machine learning methods to work individually or in groups on *transdisciplinary tasks*, to *make annotation issues explicit*, to consult and *come to agreements* using backing knowledge bases, to take advantage of *state-of-the-art machine learning methods*, and to *promote annotation via crowdsourcing*. LabelMovie can be improved in many ways. LabelMovie can be used for special machine annotations in specific application domains.

References

[1] M. Del Fabro, B. Münzer, and L. Böszörményi, "Smart video browsing with augmented navigation bars," in *Adv. in Multimedia Model.* Springer, 2013, pp. 88–98.

[2] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comp. Vis.*, vol. 101, pp. 184–204, 2013.

[3] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video annotation and retrieval using ontologies and rule learning," *IEEE MultiMedia*, vol. 17, pp. 80–88, 2010.

[4] Y.-G. Jiang, Q. Dai, J. Wang, C.-W. Ngo, X. Xue, and S.-F. Chang, "Fast semantic diffusion for large-scale context-based image and video annotation," *IEEE Tr. Image Proc.*, vol. 21, pp. 3080–3091, 2012.

[5] N. Tan and J.-C. Martin, "Review of anvil," 2011. [Online]. Available: http://www.anvil-software.org/

[6] A. Lőrincz, G. Molnár, L. A. Jeni, Z. Tősér, A. Rausch, and J. F. Cohn, "Towards entertaining and efficient educational games," NIPS Workshop on Data Driven Education, December 2013. [Online]. Available: http://lytics.stanford.edu/datadriveneducation/papers/lorinczetal.pdf

[7] D. Sonntag, "Collaborative multimodality," *KI*, vol. 26, no. 2, pp. 161–168, 2012.

[8] M. Cuturi, "Fast global alignment kernels," in *Proc. Int. Conf. Machine Learn.*, 2011, pp. 929–936.

[9] A. Lőrincz, L. A. Jeni, Z. Szabó, J. F. Cohn, and T. Kanade, "Emotional expression classification using time-series kernels," in *IEEE Comp. Vis. Pattern Rec. Workshop*. IEEE, 2013, pp. 889–895.

[10] L. A. Jeni, A. Lőrincz, T. Nagy, Z. Palotai, J. Sebők, Z. Szabó, and D. Takács, "3D shape estimation in video sequences provides high precision evaluation of facial expressions," *Image and Vis. Comput.*, vol. 30, pp. 785–795, 2012.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Tr. Int. Syst. Techn.*, vol. 2, pp. 27:1–27:27, 2011.

[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[13] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *IEEE CVPR*, 2012, pp. 3642–3649.

[14] Y. Bengio, A. Courville, and P. Vincent, "Representation learning," *IEEE Tr. PAMI*, vol. 35, pp. 1798–1828, 2013.

[15] M. T. Maybury, *Multimedia Information Extraction*. Wiley, 2012.

[16] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval," *ACM Tr. Multimedia Comput. Commun. Appl.*, vol. 2, pp. 1–19, 2006.

[17] B. Csapó, A. Lörincz, and G. Molnár, "Innovative assessment technologies in educational games designed for young students," in *Assessment in Game-Based Learning*. Springer, 2012, pp. 235–254.

[18] A. Lőrincz and B. Pintér, "Natural language processing supported transdisciplinary crowdsourcing," in *3rd EUCogIII Conf*, 2013. [Online]. Available: http://nipg.inf.elte.hu/publications/lorincz13natural.pdf

[19] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre, "Continuous au intensity estimation using localized, sparse facial feature space," in *10th IEEE Workshop Aut. Face and Gesture Rec.* IEEE, 2013, pp. 1–7.