ONLINE HUMAN ACTION LOCALISATION BASED ON APPEARANCE AND MOTION CUES

Saha S., Cuzzolin F., - Oxford Brookes University
Sapienza M., - University of Oxford
{suman.saha-2014, fabio.cuzzolin}@brookes.ac.uk, michael.sapienza@eng.ox.ac.uk

ICVSS 2015
Sicily ~ 12-18 July
International Computer Vision Summer School

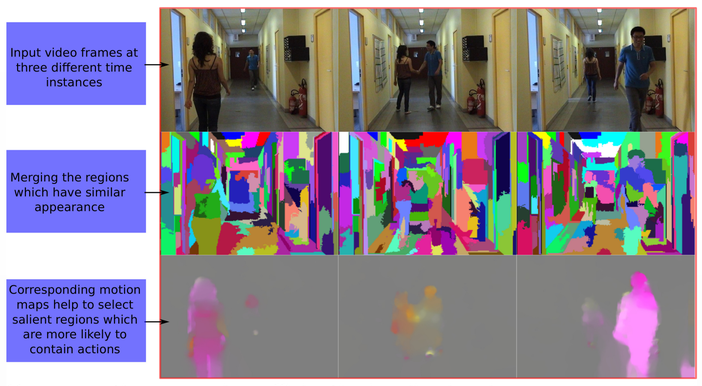
Abstract

We investigate the problem of online action localisation *in videos. Our model uses appearance and motion cues to generate region proposals from streaming video frames. Recently, deep feature representation outperforms the handcrafted features in object classification. Driven by this progress, we model our system using deep CNN features. We proposed an online incremental learning framework which initially learns from a burst of streaming video frames and iteratively updates the learner by solving a set of linear SVMs (1-vs-rest) using a batch stochastic gradient descent (SGD) algorithm with hard example mining.*

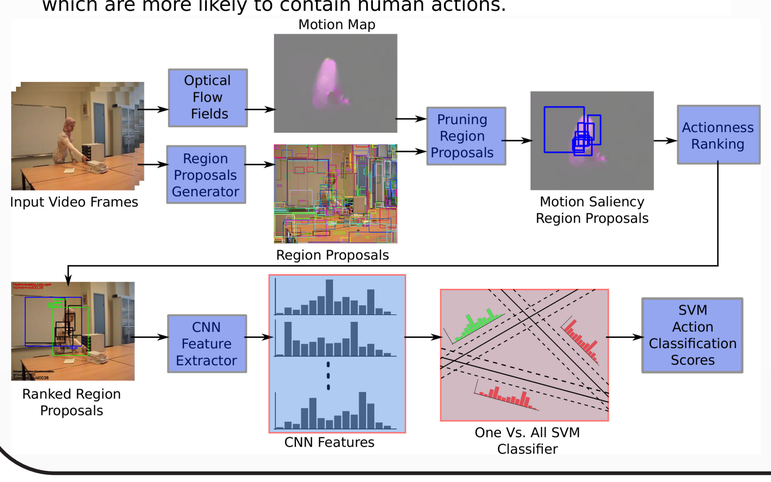
State-of-the-art

- Recently deep learning technique [1] outperforms the hand-crafted feature representation approaches in action classification [3] and detection [2].
- However, a robust **online action detection system** is yet to be addressed by the vision-community!
- Therefore, *it is worthwhile to investigate* the state-of-the-art *deep learning approach* for online action detection.

Our approach



Human actions can be efficiently localised using appearance and motion cues over space and time. Our framework generates spatio-temporal region proposals from streaming video by 1) firstly, merging the regions which have similar appearance; 2) secondly, pruning those region proposals using motion cues over consecutive video frames. Thus, motion salient space-time region proposals can be obtained which are more likely to contain human actions.



Method

1. Extract space-time *region proposals* R_i from a burst of video frames F_j using SGBH (streaming graph based hierarchical) video segmentation algorithm [7].
2. **Prune** region proposals (cf. Step-1) using **motion saliency scores** Sm obtained from *dense optical flow fields* computed over F_j [6].
3. **Rank** the motion salient region proposals (cf. Step-2) using the **intersection-over-union** (IoU) scores with respect to the ground truth annotation.
4. Obtain image patch descriptor for each ranked region proposals using a pre-trained Convolutional Neural Network (CNN) .
5. Train an *online incremental learning* algorithm with the CNN features (cf. Step-4) for action classification and detection.

Motion saliency score $S_m = \frac{\sum_{i \in R} f_m(i)}{\sum_{j \in I} f_m(j)}$; f_m is the normalised optical flow magnitude.
IoU score $a_o = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$; B_p proposed bounding box and B_{gt} ground truth annotation.

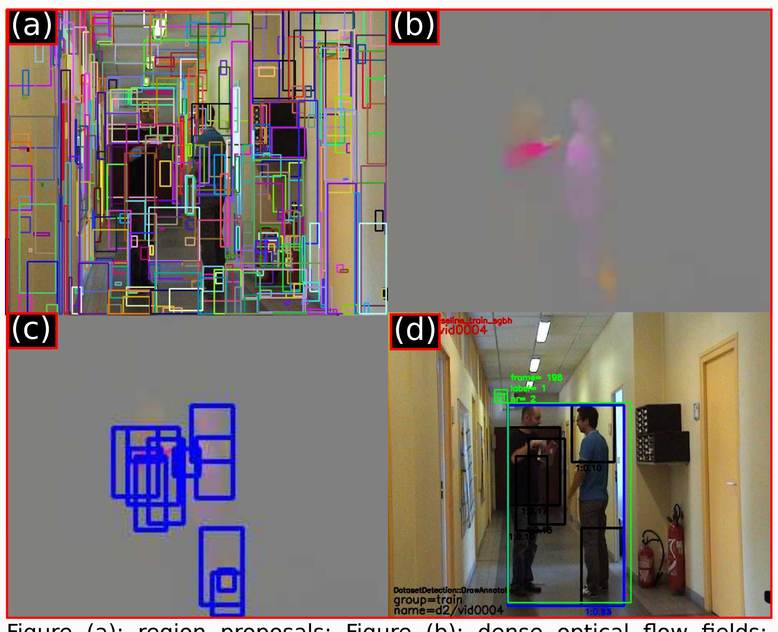
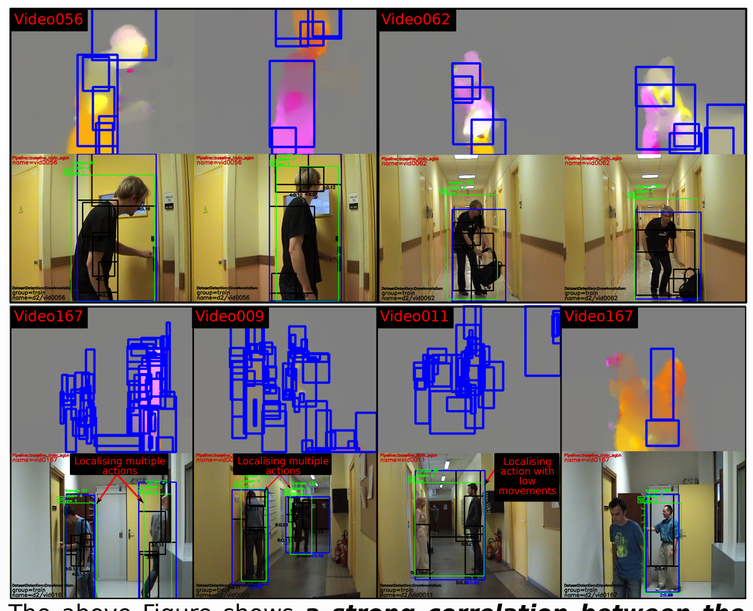


Figure (a): region proposals; Figure (b): dense optical flow fields; Figure (c): region proposals pruned using optical flow; Figure (d): region proposals finally selected by our proposed ranking algorithm.

Actionness ranking



The above Figure shows ***a strong correlation between the motion saliency measures and the IoU overlap which is necessary to ensure good performance at test time***. Notice in rows 2 and 4, our estimated actionness regions (depicted in blue bounding boxes) highly overlap with the ground truth annotations (shown in green bounding boxes). Videos contain actions such as "try enter room unsuccessfully" (video 56, 167), "leave baggage unattended" (video 62) and "put take obj into from box/ desk" (video 167) exhibit relatively higher movements, and thus, have larger mean IoU scores between 0.44 to 0.53. Whereas, actions like "typing on keyboard", "telephone conversation" and "discussion" (video 11) involve less movements and have lower IoU scores between 0.3 to 0.41.

Online learner

- Given *CNN feature vectors* $\boldsymbol{x}_i \in \mathbb{R}^n$ for region proposals R_i, a set of linear SVMs (1-vs-rest) is used to assign class labels y_i to R_i.
- In a classical SVM setting the following objective function is minimised:

$$o_{\mathcal{D}}(\hat{\boldsymbol{w}}) = \frac{1}{2}\|\hat{\boldsymbol{w}}\|^2 + C\sum_{i=1}^{D} max(0, 1 - y_i \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}_i) \quad (1)$$

where: dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_D, y_D)\}$; vector of parameters $\hat{\boldsymbol{w}} = (w, b)$; regularisation parameter C; total number of examples D; $y \in \{-1, +1\}$; $\hat{\boldsymbol{x}} = (\boldsymbol{x}, 1)$ augmented to include a bias-multiplier.

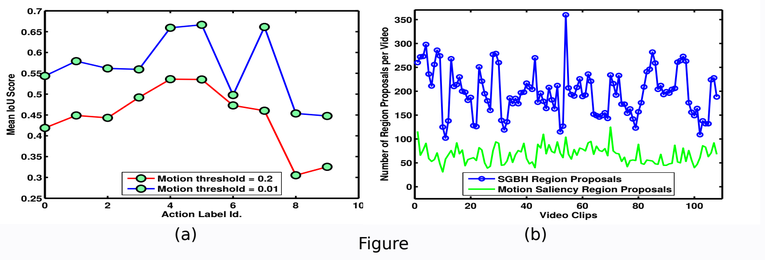
- In our case, data is streamed in time, therefore, we use a batch variant of **SGD** which iteratively updates $\hat{\boldsymbol{w}}$ by taking a step in the negative direction of the gradient w.r.t. a randomised example set $\mathcal{E}^t \subseteq \mathcal{D}$.
- Given inputs $\hat{\boldsymbol{w}}^t$, $\mathcal{H}^t$ and $\mathcal{E}^t$, the following is a single step towards minimum of the objective function in Eq. 1:

$$\hat{\boldsymbol{w}}^{t+1} := \hat{\boldsymbol{w}}^t - \alpha^t(\hat{\boldsymbol{w}}^t + C \sum_{(\hat{\boldsymbol{x}}_i, y_i) \in \mathcal{H}^t} h(\hat{\boldsymbol{w}}^t, \hat{\boldsymbol{x}}_i, y_i)) \quad (2)$$

where α^t: the *learning step size* at time t; $\mathcal{H}$: cache of hard examples [4], $\mathcal{H} := \mathcal{H} \cup sample(\mathcal{E}^t, \text{batch-size})$.

Results and Conclusion

Experimental Setup: challenging LIRIS HARL human activity dataset with 10 complex human action classes; a desktop with Intel Core i5-3570 CPU @ 3.40GHz x 4 and 32 GB RAM.



(a) (b)

Figure

Figure (a) shows the significance of the motion saliency measure over IoU score; we achieved an ***increased mean IoU of 0.6*** with a ***smaller motion threshold 0.01*** as compared to mIoU 0.44 with relatively larger threshold value 0.2; thus, ***a strong correlation between the notion actionness (or motion saliency measure) and IoU overlap can be observed***. Figure (b) shows a dramatic reduction in the image search space with **3 times** less region proposals after pruning. The IoU score is averaged over all the class-specific training videos.

Label Id.	Top 10 IoU Scores*	Label Id.	Lowest 10 IoU Scores*
3	0.8764	9	0.3192
1	0.8258	8	0.3165
7	0.8152	3	0.3018
4	0.8119	9	0.2925
4	0.8042	9	0.2907
5	0.7998	9	0.2828
4	0.7926	9	0.2718
3	0.7925	8	0.2480
7	0.7912	3	0.2453
3	0.7758	9	0.2436

Table 1

LIRIS HARL Action Class	Label Id.
discussion	0
give-object-to-person	1
put-take-obj-into-from-box-desk	2
enter-leave-room-no-unlocking	3
try-enter-room-unsuccessfully	4
unlock-enter-leave-room	5
leave-baggage-unattended	6
handshaking	7
typing-on-keyboard	8
telephone-conversation	9

Table 2

Notice in Table 1, actions 3 appears in both top-10 and lowest-10 IoU score lists. The reason for this discrepancy between IoU scores of same action classe is the SGBH region proposals.

SGBH hierarchically groups regions based on only ***a single region merging criterion***, i.e., the measure of similarity between two regions in a video is the ***Chi squared distance*** between the colour histograms of those regions in ***Lab*** colour space. Due to a single merging criterion, the space-time region proposals do not give consistent overlaps with GT annotations, also, region proposals drift over time. Thus, the resultant SGBH bounding boxes yield low IoU scores degrading the overall performance of the detection system. Table 2 shows the LIRIS HARL 10 action classes and their label ids. (*) IoU scores are averaged over each video clip.

Conclusion and future work:

- Our motion saliency algorithm is robust to detect multiple actions simultaneously.
- To improve the IoU scores, we will combine our motion saliency method with segmentation algorithms which consider a range of **diversified region merging criteria** such as "Selective Search".
- To make our motion saliency algorithm robust against multiple actions, we will use techniques like "non-maximum suppression" [5].
- We plan to **integrate motion and appearance features** by incorporating a separate CNN to encode action dynamics from multiple consecutive video frames.
- During test time we **link multiple compound hypotheses (region proposals) over individual action tubes** as per their class specific SVM-scores.

References

[1] A., Krizhevsky, S., Ilya, G., E., Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in *NIPS*, Year 2012.

[2] G., Gkioxari, J., Malik, Finding Action Tubes, in *CVPR*, Year 2015.

[3] K., Simonyan, A., Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, in *CoRR*, Year 2014.

[4] P., Felzenszwalb, R., Girshick, D., McAllester, D., Ramanan, Object detection with discriminatively trained part based models, *PAMI*, Year 2010.

[5] R., Girshick, J., Donahue, T., Darrel, J., Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *CVPR*, Year 2014.

[6] T., Brox, A., Bruhn, N., Papenberg, j., Weickert, High accuracy optical flow estimation based on a theory for warping, in Proc. *ECCV*, Year 2004.

[7] X., Chenliang, X., Caiming, J., Corso, Streaming Hierarchical Video Segmentation, in *ECCV*, Year 2012.