# Detection and Synchronization of Video Sequences for Event Reconstruction

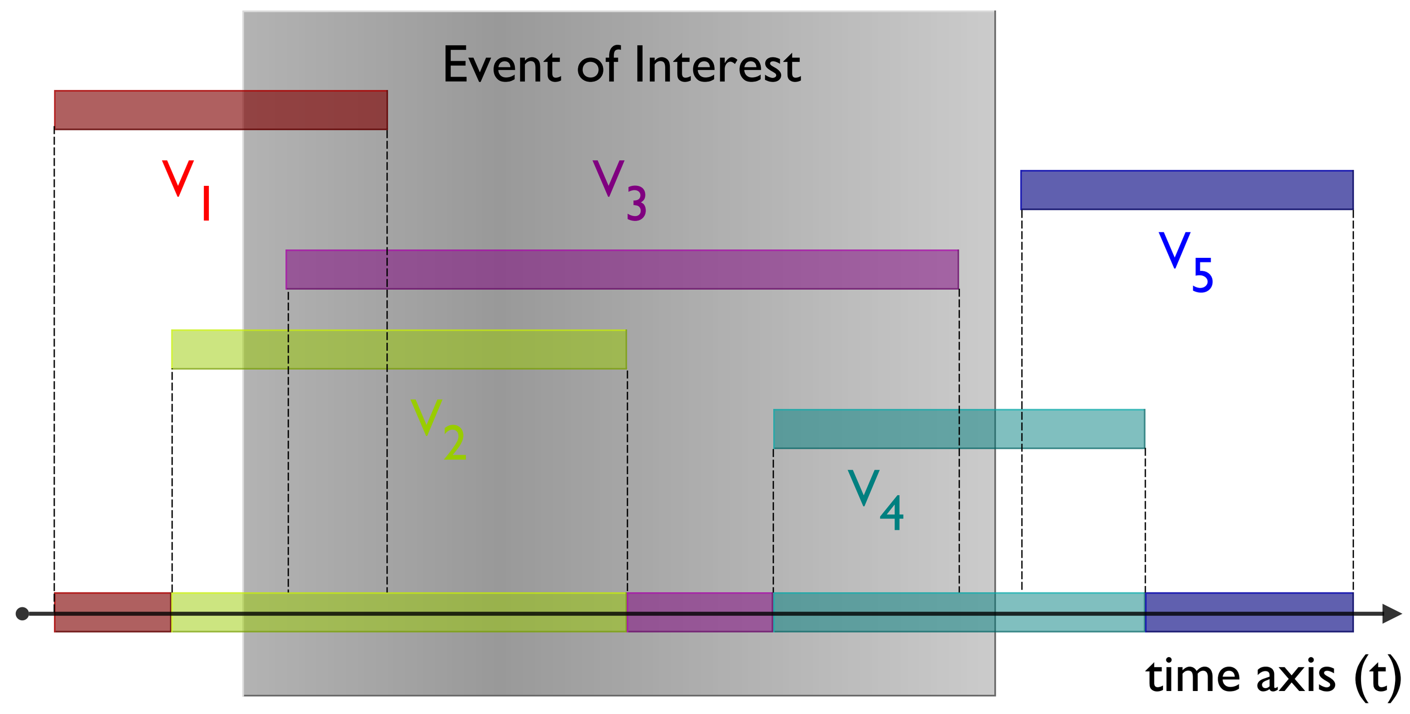
Giuliano Pinheiro[1], Marcos Cirne[1], Paolo Bestagini[2], Stefano Tubaro[2], Anderson Rocha[1]

[1]Institute of Computing, University of Campinas (UNICAMP), Campinas, SP - Brazil

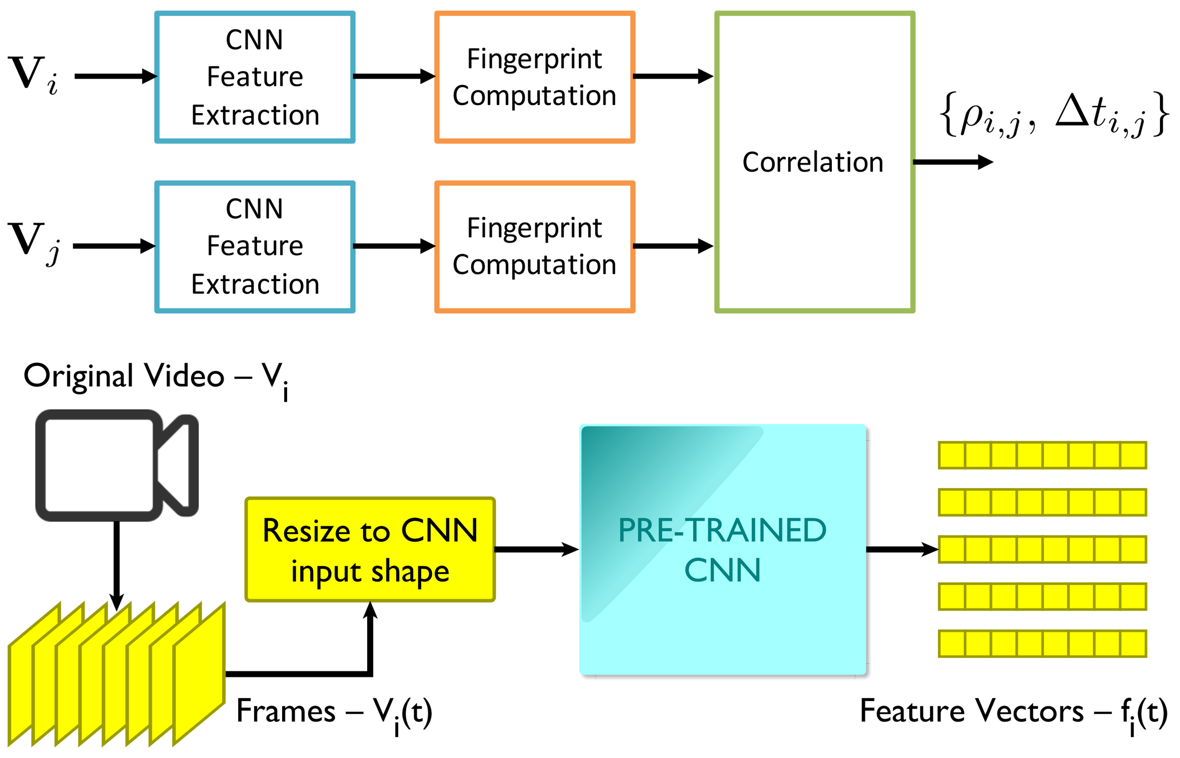
[2]Dipartimento di Elettronica, Informazione e Bioingegneria Politecnico di Milano, Milan, Italy

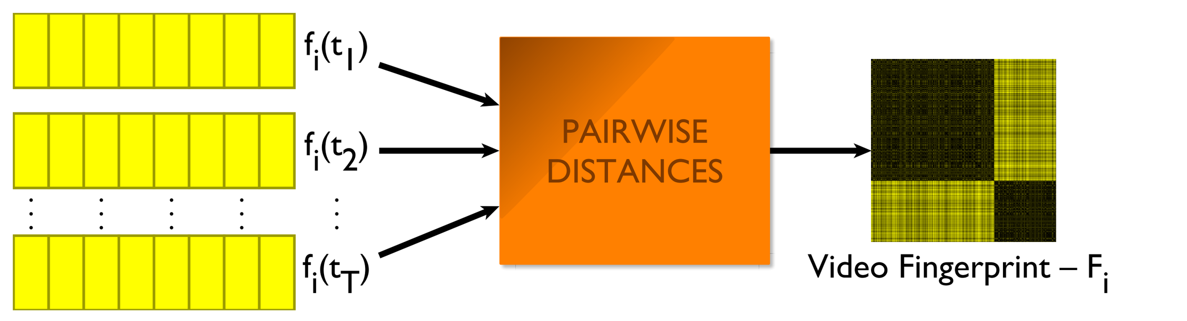
## 1. PROBLEM CHARACTERIZATION


Given an event $E$, captured by different cameras from different viewpoints with recordings at different time instants and durations, our goal is to align them spatially and temporally.




## 2. PROPOSED METHODOLOGY

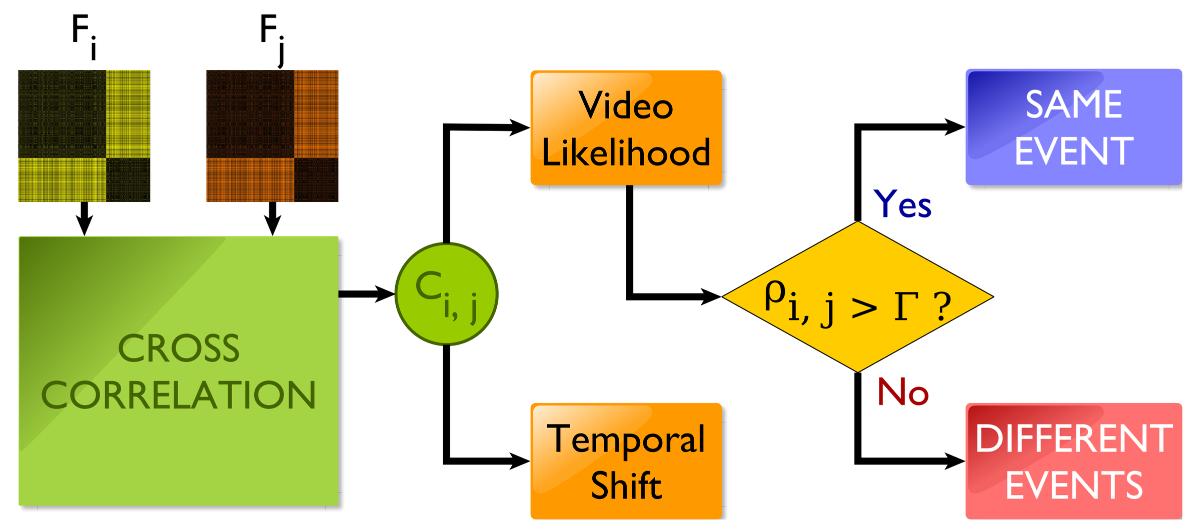



$$\mathbf{F}_i(t_a, t_b) = ||\mathbf{f}_i(t_a) - \mathbf{f}_i(t_b)||_1, \quad t_a, t_b \in [t_i, t_i + T_i],$$

Once the fingerprints $\mathbf{F}_i$ from video $\mathbf{V}_i$ and $\mathbf{F}_j$ from video $\mathbf{V}_j$ have been computed, they are compared by computing the complete normalized cross-correlation matrix.

$$\mathbf{C}_{i,j}(\tau_a, \tau_b) = \mathcal{F}^{-1}\left(\mathcal{F}\left(\frac{\mathbf{F}_i - \mu(\mathbf{F}_i)}{||\mathbf{F}_i||_2}\right) \cdot \mathcal{F}\left(\frac{\mathbf{F}_j - \mu(\mathbf{F}_j)}{||\mathbf{F}_j||_2}\right)^*\right)$$




Video Likelihood: $\rho_{i,j} = \max_{\tau_a, \tau_b}(\mathbf{C}_{i,j}(\tau_a, \tau_b))$

Temporal Shift: $\Delta t_{i,j} = \arg\max_{\tau_a, \tau_b}(\mathbf{C}_{i,j}(\tau_a, \tau_b))$

## 3. EXPERIMENTAL SETUP

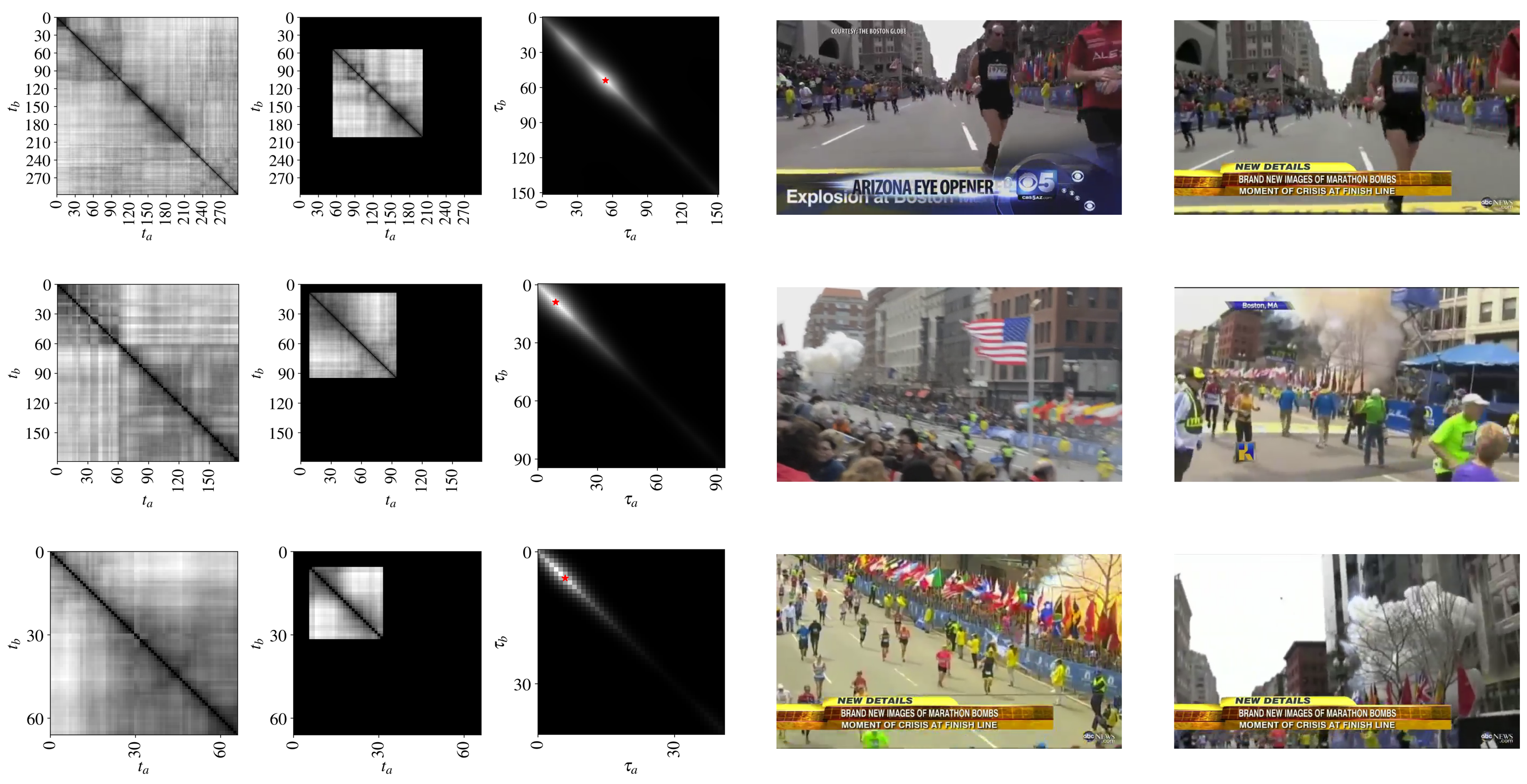
In our experiments, we investigated the use of two pre-trained CNN architectures:

- **VGG19** [1]: Input size $224 \times 224$, FC-2 layer used for feature extraction
- **Inception ResNet V2** [2]: Input size $299 \times 299$, *avg-pool* layer used for feature extraction

Also, we used two types of datasets:

- a set of 9 different YouTube videos with several viewpoints from the Boston Marathon bombing attack in 2013;
- a synthetic dataset of almost 800 edited videos coming from 19 video sequences. For each sequence, we generated a series of 42 near-duplicate videos obtained by randomly applying cropping, rotation, flipping, brightness adjustment, and contrast enhancement.

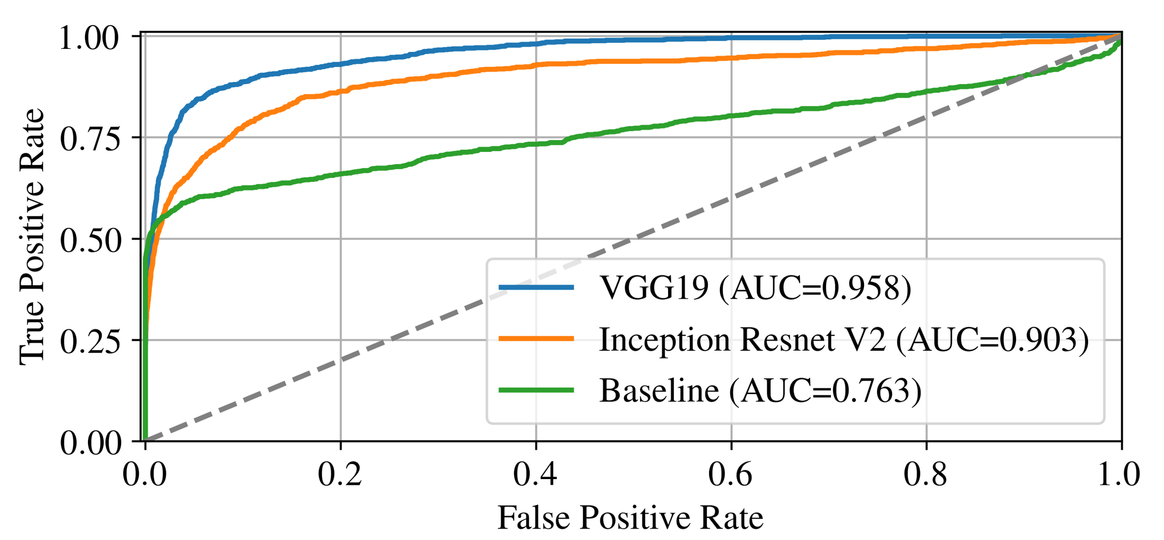
## 4. RESULTS ON REAL USE-CASE



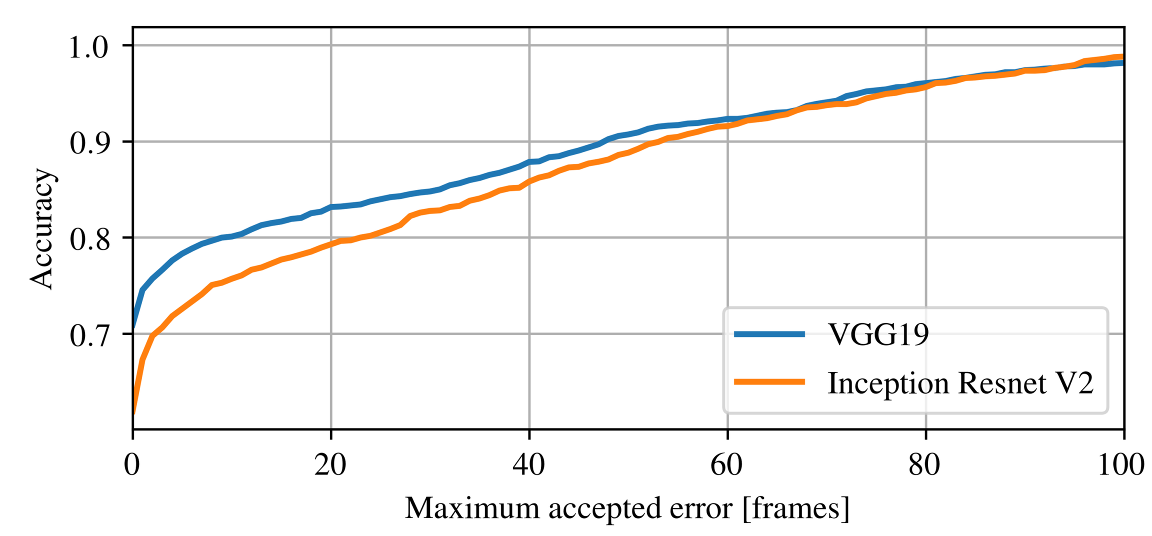
a) $\mathbf{F}_i$ b) $\mathbf{F}_j$ c) $\mathbf{C}_{i,j}$ d) $\mathbf{V}_i(\Delta t_{i,j})$ e) $\mathbf{V}_j(0)$

Three examples from Boston Marathon use-case: (a) fingerprint $\mathbf{F}_i$; (b) fingerprint $\mathbf{F}_j$ re-aligned with $\mathbf{F}_i$; (c) correlation $\mathbf{C}_{i,j}$ enabling the correct re-alignment, with maximum location highlighted with a red asterisk; (d) and (e) estimated pair of aligned frames.

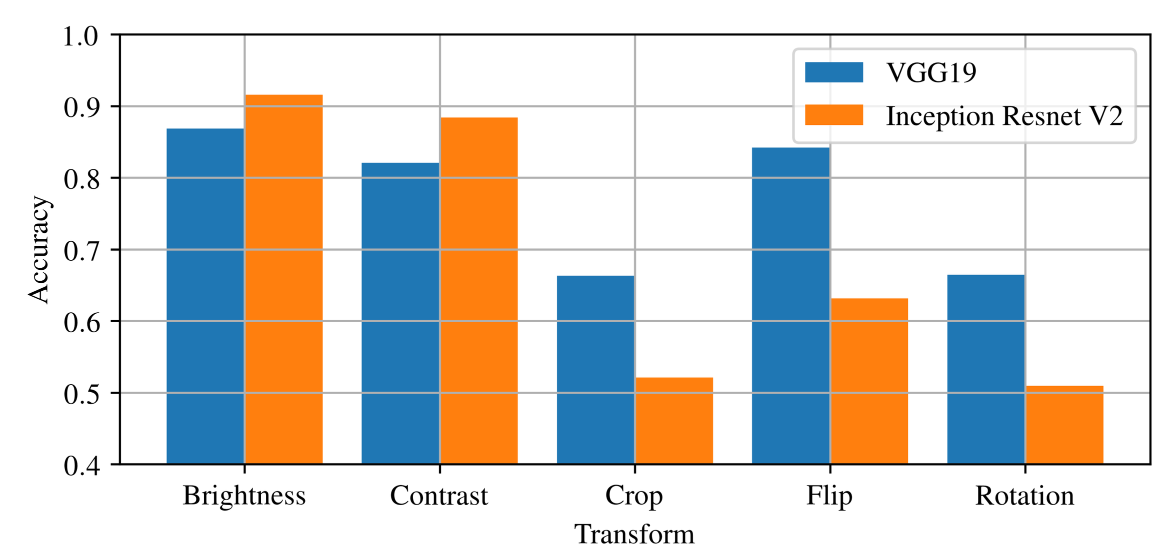
## 5. RESULTS ON SYNTHETIC DATA




ROC curves showing video detection performance using the proposed method based on different features (i.e., VGG19 and InceptionResNetV2) and the baseline solution based on [3].




Video alignment accuracy considering a maximum accepted error in frames. VGG19 provides 70% accuracy in terms of perfect alignment, and more than 90% if a maximum error of 60 frames (i.e., 2 seconds) is accepted.




Video alignment accuracy considering different features and transformations.

## 6. CONCLUSIONS AND FUTURE WORK

- Pre-trained CNNs are particularly robust and less prone to overfitting for this problem
- Good results on synthetic video data motivated us to test the approach on a real-world use case with promising results
- Future work will be devoted to the use of 3D feature vectors that capture the temporal evolution of the scene, rather than working on a frame-by-frame basis.

## 7. REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[2] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016.

[3] S. Lameri, P. Bestagini, A. Melloni, S. Milani, A. Rocha, M. Tagliasacchi, and S. Tubaro, "Who is my parent? Reconstructing video sequences from partially matching shots," in *IEEE International Conference on Image Processing (ICIP)*, 2014.

## 8. ACKNOWLEDGEMENTS

We thank the financial support of São Paulo Research Foundation (FAPESP) through grant #2017/12646-3, DéjàVu project. Finally, the authors also thank Nvidia Corporation for GPUs donated through the Nvidia GPU grant program.