Table 2: Results of asynchronous fusion network (split 1)

Methods	UCF101	HMDB51
Two-stream baseline	89.8%	58.4%
Baseline+SYN	89.7%	–
Baseline+ASYN ($\Delta = 1$)	90.3%	–
Baseline+ASYN ($\Delta = 5$)	**91.0%**	**60.9%**
CO2FI	92.8%	67.9%
CO2FI+ASYN ($\Delta = 5$)	**93.7%**	**69.5%**

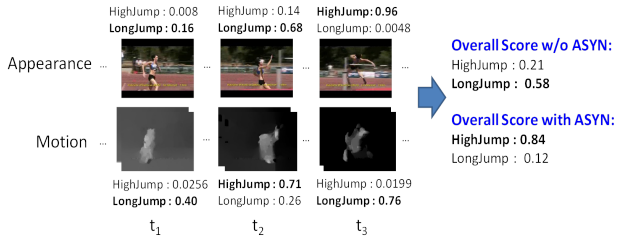

Figure 6: An example of the effect of the asynchronous fusion network: since two streams create high prediction scores for the ground-truth class "Highjump" at different time points, the recognition result will be easily confused if not considering the stream-wise asynchronous pattern.

vious since the longer-term asynchronous patterns are not properly captured. Comparatively, when fusing stream-wise features with larger temporal distances (*Baseline+ASYN* ($\Delta = 5$)), we can obtain more noticeable improvements. This demonstrates that the asynchrony between different information streams indeed affects action recognition performances. Moreover, from the lower part of Table 2, we can also observe that when combining our asynchronous fusion network with the coarse-to-fine network, we can obtain further improved recognition performances by leveraging both mutli-granularity features and stream-wise complementary information (*CO2FI+ASYN* ($\Delta = 5$)).

Fig. 6 further shows an example about the effect of the asynchronous fusion network. In Fig. 6, since the two information streams of the "Highjump" video have asynchronous patterns, they create high prediction scores for the ground-truth action class at different time points (e.g., t_3 for appearance stream and t_2 for motion stream in Fig. 6). If we simply sum up the prediction scores over time or only consider the stream-wise correlation at the same time, the final recognition result will be confused with other action classes (cf. *Overall score w/o ASYN*). Comparatively, if we consider the asynchrony between streams and allow stream-wise feature fusion at different time, the complementary information between streams can be more properly used, resulting in a correct result (cf. *Overall score with ASYN* in Fig. 6).

6.4 Comparison with the state-of-the-art

Table 3 compares our approach (*CO2FI+ASYN*) with the state-of-the-art methods. Since many works reported results by performing a late fusion with hand-crafted IDT features (**?**), we also show fusion result of our approach (*CO2FI+ASYN+IDT*). Note that in this experiment, we adopt three training/testing splits on both datasets in order to have a fair comparison with other methods.

Table 3: Comparison of different methods (3 splits)

Methods	UCF101	HMDB51
C3D (3 nets) [Tran *et al.* 2015]	85.2%	–
AdaScan [Kar *et al.* 2017]	89.4%	54.9%
TDD+FV [Wang *et al.* 2015]	90.3%	63.2%
GRP [Cherian *et al.* 2017]	91.9%	65.4%
Three-stream sDTD [Shi *et al.* 2017]	92.2%	65.2%
Transformations [Wang *et al.* 2016]	92.4%	62.0%
Two-Stream Fusion [Feichtenhofer *et al.* 2016]	92.5%	65.4%
KVMF [Zhu *et al.* 2016]	93.1%	63.3%
ST-ResNet [Feichtenhofer *et al.* 2016]	93.4%	66.4%
L^2STM [Sun *et al.* 2017]	93.6%	66.2%
ST-VLMPF [Duta *et al.* 2017]	93.6%	**69.5%**
TSN (2 modelities) [Wang *et al.* 2016b]	94.0%	68.5%
CO2FI + ASYN	**94.3%**	69.0%
Dynamic Image Networks + IDT [Bilen *et al.* 2016]	89.1%	65.2%
AdaScan + IDT [Kar *et al.* 2017]	91.3%	61.0%
TDD + IDT [Wang *et al.* 2015]	91.5%	65.9%
GRP + IDT [Cherian *et al.* 2017]	92.3%	67.0%
ST-ResNet + IDT [Feichtenhofer *et al.* 2016]	94.6%	70.3%
CO2FI + ASYN + IDT	**95.2%**	**72.6%**

From Table 3, we can see that our approach has better performances than most of the state-of-the-art methods. Specifically, when comparing with the most recent works using ResNet (*ST-ResNet*) or introducing an additional information stream (*ST-VLMPF*), our approach can also obtain similar or better results. This demonstrates the effectiveness of our proposed approach. Note that comparing with *ST-ResNet* and *ST-VLMPF*, we use a relatively short ConvNet (VGG-16) and do not introduce additional information streams. It is expected that the performances of our approach can be further improved if using deeper ConvNets such as ResNet or including more information streams. Moreover, our approach fused with IDT (*CO2FI+ASYN+IDT*) also performs better than other IDT-fused methods. This furhter indicates the robustness of our approach in improving performances.

7 Conclusion

This paper presents a novel framework for action recognition. Our framework consists of two key ingredients: 1) a coarse-to-fine network, which extracts and integrates deep features from multiple action class granularities to obtain a more precise feature representation for actions; 2) an asynchronous fusion network which integrates stream-wise features at different time points for better leveraging the information in multiple streams. Experimental results show that our approach achieves the state-of-the-art performance.

Adipisci minus id dicta in, dolorum cum vel dolores doloremque aut distinctio nisi non.Consequuntur soluta molestiae sunt eius perspiciatis iure aliquam numquam, placeat aliquid molestias?Porro atque necessitatibus dolorem, dolor maxime explicabo at reiciendis, vitae illum odio voluptates minima.Velit suscipit ab maxime, totam ex animi necessitatibus consequatur, iusto atque perferendis neque porro asperiores doloremque quia dolores explicabo, nostrum architecto voluptatibus exercitationem eveniet doloremque repre-

henderit quibusdam culpa.Corporis non aspernatur magnam quas, aperiam sed soluta?Distinctio ipsam pariatur dolores officiis repellendus sit, minima