Weakly-Supervised Temporal Action Alignment Driven by Unbalanced Spectral Fused Gromov-Wasserstein Distance

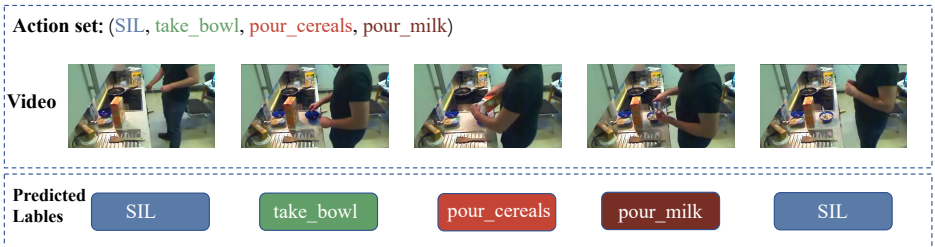
Dixin Luo[1] Yutong Wang[1] Angxiao Yue[1] Hongteng Xu[2]

[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]Gaoling School of Artificial Intelligence, Renmin University of China

August 16, 2022

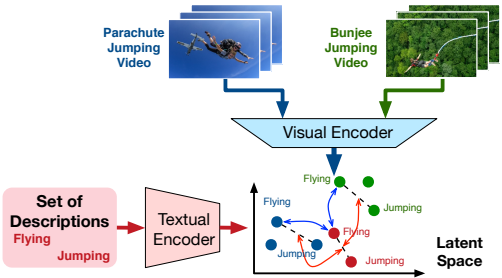
Background

- Weakly supervised temporal action alignment.
  - Task: Assign each video with a set of (unsorted) texts and aim at matching each frame with a text in the set.
  - Challenges: Correspondence between the frames and the texts is unknown and lack of text order information leads to sub-optimal performance.



Motivation

- The perspective of computational optimal transport.

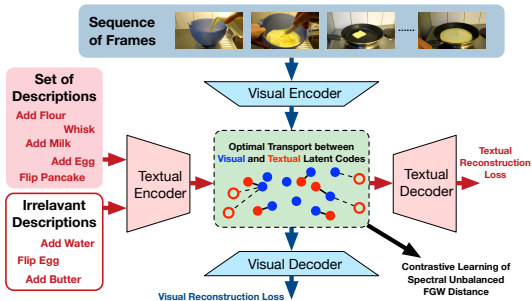


- Wasserstein distance captures point-wise similarity while Gromov-Wasserstein (GW) Distance focus on pair-wise comparison.
- Fused Gromov-Wasserstein Distance (FGW) combines both of them.

$$d_{\text{fgw}}(\boldsymbol{V}, \boldsymbol{W}; \beta) = \min_{\boldsymbol{T} \in \Pi(\boldsymbol{u}, \boldsymbol{\mu})} \underbrace{(1-\beta)\langle \boldsymbol{D}_{vw},\ \boldsymbol{T}\rangle}_{\text{Wasserstein term}} + \underbrace{\beta\langle -\boldsymbol{D}_v \boldsymbol{T} \boldsymbol{D}_w^T,\ \boldsymbol{T}\rangle}_{\text{GW term}}.$$

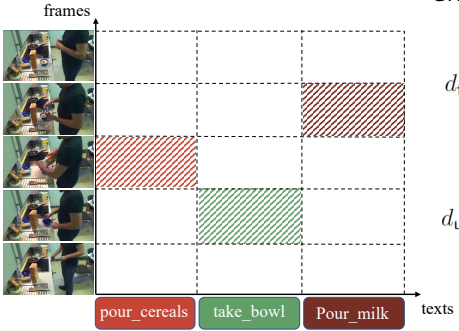
Our Approach: Overview

- Unbalanced spectral fused Gromov-Wasserstein (US-FGW) distance is proposed to capture the correspondence between frames and textual descriptions in their latent space
- A new contrastive learning framework is applied for the unified learning of the visual and textual auto-encoders



Our Approach

- Unbalanced spectral fused Gromov-Wasserstein (US-FGW) distance
  - Filter out meaningless background frames - Unbalanced setting
  - Distinguish the high-dimensional latent codes - Spectral setting



Given frames $\boldsymbol{V}$, text words $\boldsymbol{W}$:

$$d_{\text{fgw}}(\boldsymbol{V},\boldsymbol{W};\beta) = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{u},\boldsymbol{\mu})} \underbrace{(1-\beta)\langle \boldsymbol{D}_{vw},\ \boldsymbol{T}\rangle}_{\text{Wasserstein term}} + \underbrace{\beta\langle -\boldsymbol{D}_v\boldsymbol{T}\boldsymbol{D}_w^T,\ \boldsymbol{T}\rangle}_{\text{GW term}}.$$

$$\begin{aligned} d_{\text{us-fgw}}(\boldsymbol{V},\boldsymbol{W};\beta,\tau) = \min_{\boldsymbol{T}} \ & (1-\beta)\langle -\boldsymbol{K}_{vw},\boldsymbol{T}\rangle + \beta\langle -\boldsymbol{K}_v\boldsymbol{T}\boldsymbol{K}_w^T,\boldsymbol{T}\rangle \\ & + \tau\Big(\mathsf{KL}\big(\boldsymbol{T}\mathbf{1}_J \| \frac{1}{I}\mathbf{1}_I\big) + \mathsf{KL}\big(\boldsymbol{T}^T\mathbf{1}_I \| \frac{1}{J}\mathbf{1}_J\big)\Big). \end{aligned}$$

Our Approach

- Leverage the **Bregman ADMM algorithm** to compute the US-FGW distance
  1. Rewrite $d_{\text{us-fgw}}$ by introducing three auxiliary variables $\boldsymbol{S}$, $\boldsymbol{u}$ and $\boldsymbol{\mu}$:

$$\begin{aligned} &\min_{\boldsymbol{T},\boldsymbol{S},\boldsymbol{u},\boldsymbol{\mu}} (1-\beta)\langle -\boldsymbol{K}_{vw}, \boldsymbol{T}\rangle + \beta\langle -\boldsymbol{K}_v \boldsymbol{S} \boldsymbol{K}_w^T, \boldsymbol{T}\rangle + \\ &\quad \tau\Big(\mathsf{KL}(\boldsymbol{u}\|\frac{1}{I}\mathbf{1}_I) + \mathsf{KL}(\boldsymbol{\mu}\|\frac{1}{J}\mathbf{1}_J)\Big) \\ &s.t.\ \boldsymbol{T} = \boldsymbol{S},\ \boldsymbol{T}\mathbf{1}_J = \boldsymbol{u},\ \boldsymbol{S}^T\mathbf{1}_I = \boldsymbol{\mu}. \end{aligned} \tag{1}$$

  2. Rewrite this formula in a Bregman-augmented Lagrangian form by introducing three dual variables $\boldsymbol{Z}$, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$.
  3. Update the primal, the auxiliary, and the dual variables iteratively till the variable $\boldsymbol{T}$ converges.

Our Approach

- Leverage the **Bregman ADMM algorithm** to compute the US-FGW distance
  - 3.1 Rewrite (1) in the following the Bregman-augmented Lagrangian form for $\boldsymbol{T}$ and update $\boldsymbol{T}$ in a closed form:

$$\begin{aligned}\boldsymbol{T}^{(k+1)} &= \arg\min_{\boldsymbol{T}\in\Pi(\boldsymbol{u}^{(k)},\cdot)}(\beta-1)\langle \boldsymbol{K}_{vw},\boldsymbol{T}\rangle - \beta\langle \boldsymbol{K}_v\boldsymbol{S}^{(k)}\boldsymbol{K}_w^T,\boldsymbol{T}\rangle \\ &\quad + \langle \boldsymbol{Z}^{(k)},\boldsymbol{T}-\boldsymbol{S}^{(k)}\rangle + \rho\mathsf{KL}(\boldsymbol{T}\|\boldsymbol{S}^{(k)}) \\ &= \mathsf{diag}(\boldsymbol{u}^{(k)})\sigma_{\mathsf{r}}\left(\frac{(1-\beta)\boldsymbol{K}_{vw}+\beta\boldsymbol{K}_v\boldsymbol{S}^{(k)}\boldsymbol{K}_w^T-\boldsymbol{Z}^{(k)}}{\rho}+\log\boldsymbol{S}^{(k)}\right),\end{aligned} \tag{2}$$

  - 3.2 Similarly, we can consider the Bregman-augmented Lagrangian form for $\boldsymbol{S}$, $\boldsymbol{u}$, $\boldsymbol{\mu}$:

$$\begin{aligned}\boldsymbol{S}^{(k+1)} &= \arg\min_{\boldsymbol{S}\in\Pi(\cdot,\boldsymbol{\mu}^{(k)})} -\beta\langle \boldsymbol{K}_v^T\boldsymbol{T}^{(k+1)}\boldsymbol{K}_w,\boldsymbol{S}\rangle \\ &\quad + \langle \boldsymbol{Z}^{(k)},\boldsymbol{T}^{(k+1)}-\boldsymbol{S}\rangle + \rho\mathsf{KL}(\boldsymbol{S}\|\boldsymbol{T}^{(k+1)}) \\ &= \sigma_{\mathsf{c}}\left(\frac{\beta\boldsymbol{K}_v^T\boldsymbol{T}^{(k+1)}\boldsymbol{K}_w+\boldsymbol{Z}^{(k)}}{\rho}+\log\boldsymbol{T}^{(k+1)}\right)\mathsf{diag}(\boldsymbol{\mu}^{(k)}),\end{aligned} \tag{3}$$

Our Approach

- Leverage the **Bregman ADMM algorithm** to compute the US-FGW distance

$$\begin{aligned} &\min_{\boldsymbol{u}\in\Delta^{I-1}} \tau \mathsf{KL}(\boldsymbol{u}\|\frac{1}{I}\mathbf{1}_I) + \langle \boldsymbol{z_1}^{(k)}, \boldsymbol{u}\rangle + \rho \mathsf{KL}(\boldsymbol{u}\|\boldsymbol{T}^{(k+1)}\mathbf{1}_J) \\ \Rightarrow\quad &\boldsymbol{u}^{(k+1)} = \sigma\left(\frac{\rho\log(\boldsymbol{T}^{(k+1)}\mathbf{1}_J) + \tau\log\frac{1}{I}\mathbf{1}_I - \boldsymbol{z_1}^{(k)}}{\rho+\tau}\right), \end{aligned} \tag{4}$$

$$\begin{aligned} &\min_{\boldsymbol{\mu}\in\Delta^{J-1}} \tau \mathsf{KL}(\boldsymbol{\mu}\|\frac{1}{J}\mathbf{1}_J) + \langle \boldsymbol{z_2}^{(k)}, \boldsymbol{\mu}\rangle + \rho \mathsf{KL}(\boldsymbol{\mu}\|(\boldsymbol{S}^{(k+1)})^T\mathbf{1}_I) \\ \Rightarrow\quad &\boldsymbol{\mu}^{(k+1)} = \sigma\left(\frac{\rho\log((\boldsymbol{S}^{(k+1)})^T\mathbf{1}_I) + \tau\log\frac{1}{J}\mathbf{1}_J - \boldsymbol{z_2}^{(k)}}{\rho+\tau}\right), \end{aligned} \tag{5}$$

3.3 Update the dual variables via the ADMM manner:

$$\begin{aligned} \boldsymbol{Z}^{(k+1)} &= \boldsymbol{Z}^{(k)} + \rho(\boldsymbol{T}^{(k+1)} - \boldsymbol{S}^{(k+1)}), \\ \boldsymbol{z_1}^{(k+1)} &= \boldsymbol{z_1}^{(k)} + \rho(\boldsymbol{u}^{(k+1)} - \boldsymbol{T}^{(k+1)}\mathbf{1}_J), \\ \boldsymbol{z_2}^{(k+1)} &= \boldsymbol{z_2}^{(k)} + \rho(\boldsymbol{\mu}^{(k+1)} - (\boldsymbol{S}^{(k+1)})^T\mathbf{1}_I). \end{aligned} \tag{6}$$

Our Approach

- A new contrastive learning framework
  - **Generalization of classic contrastive learning.** Given a set of frames, maximize the difference between the conditional distribution of positive texts and that of negative texts, i.e.,

$$\max \mathbb{E}_{\mathcal{V}\sim p_{\mathcal{D}}}[d(p_{\mathcal{P}|\mathcal{V}}, p_{\mathcal{N}|\mathcal{V}})],$$

  - **The proposed US-FGW contrastive learning.** Take US-FGW distance as the matric d, and relax the above objective funtion, i.e.,

$$\begin{aligned}&\mathbb{E}_{\mathcal{V}\sim p_{\mathcal{D}}}[d_{\text{us-fgw}}(p_{\mathcal{P}|\mathcal{V}}, p_{\mathcal{N}|\mathcal{V}})]\\ \geq&\mathbb{E}_{\mathcal{V}\sim p_{\mathcal{D}}}[d_{\text{us-fgw}}(p_{\mathcal{N}|\mathcal{V}}, p_{\mathcal{V}}) - d_{\text{us-fgw}}(p_{\mathcal{P}|\mathcal{V}}, p_{\mathcal{V}})],\end{aligned}$$

$p_{\mathcal{D}}$: the distribution of training data d: the (pseudo) metric of the distributions

$p_{\mathcal{P}|\mathcal{V}}, p_{\mathcal{N}|\mathcal{V}}$: the positive and the negative text distributions conditioned on the frame set $\mathcal{V}$

Learning Strategy

- The overall learning strategy is as follows:

$$\min_{\underbrace{f_v, g_v, f_w, g_w}_{\text{Auto-Encoders}}} \Sigma_{(\mathcal{V}_n, \mathcal{W}_n, \mathcal{W}'_n) \in \mathcal{D}} \Big(\underbrace{\ell_v(\mathcal{V}_n, g_v(f_v(\mathcal{V}_n)))}_{\text{Reconstruction loss of frames}}$$
$$+ \underbrace{\ell_w(\mathcal{W}_n, g_w(f_w(\mathcal{W}_n)))}_{\text{Reconstruction loss of words}} + \gamma \big(\underbrace{d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n); \beta, \tau)}_{\text{Positive US-FGW distance}}$$
$$- \underbrace{d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}'_n)}_{\text{Negative US-FGW distance}}; \beta, \tau) \big) \Big).$$

Nested alternating optimization

Input : Data $\mathcal{D} = \{(\mathcal{V}_n, \mathcal{W}_n)\}_{n=1}^{N}$, a vocabulary set $\mathcal{W}_{all}$, and hyper-parameters β, τ, γ

Output : Visual auto-encoder f_v, g_v, textual auto-encoder f_w, g_w

1: **for** $m = 0, ..., M-1$ (Outer Loop) **do**
2: Sample $\{\mathcal{V}_n, \mathcal{W}_n\}$ randomly from $\mathcal{D}$.

Outer loop:
Compute US-FGW distances and update the autoencoders iteratively.

Inner loop:
Update the optimal matrics when computing the US-FGW distances.

Nested alternating optimization

Input : Data $\mathcal{D} = \{(\mathcal{V}_n, \mathcal{W}_n)\}_{n=1}^{N}$, a vocabulary set $\mathcal{W}_{all}$, and hyper-parameters β, τ, γ

Output : Visual auto-encoder f_v, g_v, textual auto-encoder f_w, g_w

1: **for** $m = 0, ..., M-1$ (Outer Loop) **do**
2: Sample $\{\mathcal{V}_n, \mathcal{W}_n\}$ randomly from $\mathcal{D}$.
3: Construct a negative set $\mathcal{W}_n'$ randomly from $\mathcal{W}_{\text{all}} \setminus \mathcal{W}_n$.

Outer loop:
Compute US-FGW distances and update the autoencoders iteratively.

Inner loop:
Update the optimal matrics when computing the US-FGW distances.

Nested alternating optimization

Input : Data $\mathcal{D} = \{(\mathcal{V}_n, \mathcal{W}_n)\}_{n=1}^{N}$, a vocabulary set $\mathcal{W}_{all}$, and hyper-parameters β, τ, γ

Output : Visual auto-encoder f_v, g_v, textual auto-encoder f_w, g_w

1: **for** $m = 0, ..., M-1$ (Outer Loop) **do**
2: Sample $\{\mathcal{V}_n, \mathcal{W}_n\}$ randomly from $\mathcal{D}$.
3: Construct a negative set $\mathcal{W}_n'$ randomly from $\mathcal{W}_{\text{all}} \setminus \mathcal{W}_n$.
4: **Compute US-FGW distances** (Inner Loop):
5: Obtain $T_n^+ \leftarrow d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n))$ via B-ADMM.
6: Obtain $T_n^- \leftarrow d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n'))$ via B-ADMM.

Outer loop:
Compute US-FGW distances and update the autoencoders iteratively.

Inner loop:
Update the optimal matrics when computing the US-FGW distances.

Nested alternating optimization

Input : Data $\mathcal{D} = \{(\mathcal{V}_n, \mathcal{W}_n)\}_{n=1}^{N}$, a vocabulary set $\mathcal{W}_{all}$, and hyper-parameters β, τ, γ

Output : Visual auto-encoder f_v, g_v, textual auto-encoder f_w, g_w

Outer loop:
Compute US-FGW distances and update the autoencoders iteratively.

Inner loop:
Update the optimal matrics when computing the US-FGW distances.

1: **for** $m = 0, ..., M-1$ (Outer Loop) **do**
2: Sample $\{\mathcal{V}_n, \mathcal{W}_n\}$ randomly from $\mathcal{D}$.
3: Construct a negative set $\mathcal{W}'_n$ randomly from $\mathcal{W}_{\text{all}} \setminus \mathcal{W}_n$.
4: **Compute US-FGW distances** (Inner Loop):
5: Obtain $T_n^+ \leftarrow d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n))$ via B-ADMM.
6: Obtain $T_n^- \leftarrow d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}'_n))$ via B-ADMM.
7: **Update autoencoders:**
8: Update f_v, f_w, g_v, g_w via Adam.
9: **end for**

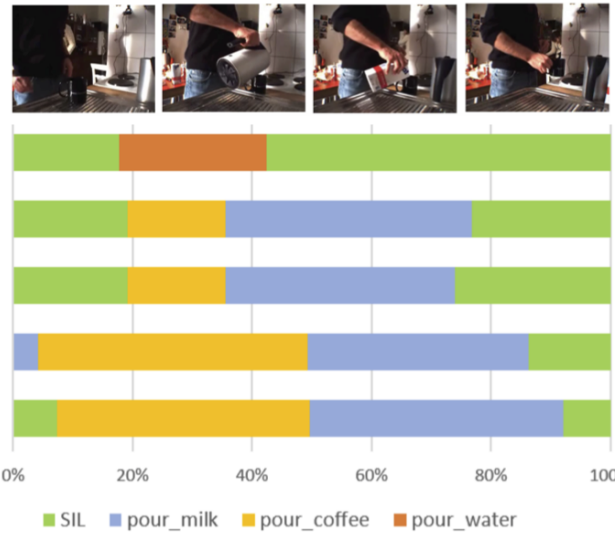
$$\min_{f_v, g_v, f_w, g_w} \sum_{(\mathcal{V}_n, \mathcal{W}_n, \mathcal{W}'_n) \in \mathcal{D}} \Big(\ell_v(\mathcal{V}_n, g_v(f_v(\mathcal{V}_n))) + \ell_w(\mathcal{W}_n, g_w(f_w(\mathcal{W}_n))) + \gamma \big(d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}_n); \beta, \tau) - d_{\text{us-fgw}}(f_v(\mathcal{V}_n), f_w(\mathcal{W}'_n); \beta, \tau) \big) \Big).$$

Quantitative Results

Table 1: Comparisons for various methods on three datasets. The baselines are categorized according to the level of supervision. For the methods with "∗", their source codes are unavailable so that we quote their results from the references, and "-" means that the corresponding results are not provided by the references.

Methods and Categories		Breakfast				Hollywood Extended				CrossTask			
		Mof	Mof-bg	IoU	IoD	Mof	Mof-bg	IoU	IoD	Mof	Mof-bg	IoU	IoD
Transcript-Supervised	$ISBA_{ED\text{-}TCN}$[CVPR 2018]	0.4548	0.2653	0.2790	0.4775	0.5553	0.4865	0.2228	0.3878	0.5174	0.5645	0.1869	0.2722
	$ISBA_{TCFPN}$[CVPR 2018]	0.4825	0.2560	0.2911	0.4972	0.5598	0.4826	0.2348	0.3958	0.5308	0.5739	0.1914	0.2706
	NNV[CVPR 2018]	0.5404	0.5131	0.4415	0.5998	0.5656	0.4737	0.3195	0.4692	0.4132	0.1991	0.1154	0.1888
	CDFL[CVPR 2019]	0.5940	0.5692	0.4391	0.6041	0.5982	0.6200	0.3514	0.5009	0.4248	0.2039	0.1256	0.2006
	TASL[ICCV 2021]	0.6042	0.5842	0.4880	0.6405	0.6066	0.5551	0.3546	0.4959	0.5148	0.4130	0.1859	0.2841
Set-Supervised	Actionset[CVPR 2018]	0.2137	0.1614	0.0504	0.1272	0.3743	0.2111	0.0966	0.1833	0.2993	0.1144	0.0268	0.0476
	SCT[CVPR 2020]	0.1220	0.1300	0.0360	0.0780	0.0550	0.2200	0.0080	0.0260	0.0640	0.1500	0.0210	0.0490
	*SCT[CVPR 2020]	0.2660	-	-	-	-	-	-	0.1770	-	-	-	-
	*SCV[CVPR 2020]	0.3020	-	-	-	-	-	-	0.1770	-	-	-	-
	*ACV[CVPR 2021]	0.3340	-	-	-	-	-	-	0.2090	-	-	-	-
	UM[AAAI 2021]	0.0383	0.0104	0.0315	0.0649	0.4053	0.2427	0.1845	0.2945	0.2053	0.1855	0.0437	0.1498
	Proposed US-FGW	0.3357	0.3577	0.1160	0.1530	0.3840	0.4082	0.2307	0.4001	0.1853	0.1907	0.0720	0.1929

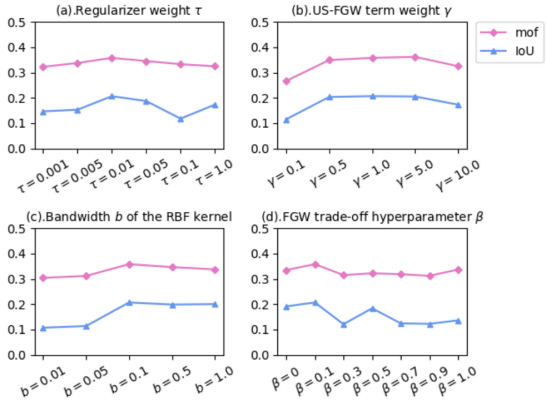
Qualitative Results



Ablation Study

Table 2: Ablation study of our US-FGW method on the Hollywood Extended dataset.

Method	Mof	Mof-bg	IoU	IoD
Solving FGW	0.3422	0.3685	0.1957	0.3666
Solving US-FGW w/o contrastive	0.3732	0.4067	0.2101	0.3780
Solving US-FGW w/ contrastive	**0.3840**	**0.4082**	**0.2307**	**0.4001**



Conclusion

- A novel optimal transport-based solution to set-supervised temporal action alignment.
- A new contrastive learning paradigm based on proposed US-FGW distance.
- An algorithm for efficient calculation of US-FGW distance leveraging the Bregman ADMM algorithm.

Thanks! Q & A

dixin.luo@bit.edu.cn
https://dixinluo.github.io