CROSS-INTERACTION-BASED MULTIMODAL FEATURE COMPARISON FOR MOVING OBJECT IDENTIFICATION IN CROWDED VIDEO SCENES

Shohruh Begmatov
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Doctor of Philosophy (PhD) in Technical Sciences, Doctoral (DSc) student
Email: sh.begmatov@tuit.uz
ORCID: 0000-0002-2441-916X
Mukhriddin Arabboev
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Doctor of Philosophy (PhD) in Technical Sciences, Doctoral (DSc) student
Email: mukhriddin.arabboev@tuit.uz
ORCID: 0000-0001-5733-5889
Akhram Nishanov
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Doctor of Science in Technical Sciences, Professor
Email: a.nishanov@tuit.uz
ORCID: 0000-0002-5652-8977
https://doi.org/10.5281/zenodo.20341226

Abstract

Identifying moving objects in crowded video scenes is difficult because appearance information alone may be unreliable. Different people or objects may have similar visual appearances, while the same object may appear differently due to pose variation, scale changes, partial occlusion, illumination variation, or low visibility. To address this problem, this paper presents a cross-interaction-based multimodal feature comparison method for moving object identification. The proposed method represents each moving object using several complementary modalities, including appearance, geometry, spatial position, context, reliability, and clothing-color features. These heterogeneous features are projected into a common latent space before comparison. For two candidate detections, modality-wise comparison features are constructed using element-wise multiplication and absolute difference. Then, a cross-interaction function learns relationships between modalities, and an MLP estimates the final similarity probability. The proposed method is especially useful in difficult cases such as occlusion, lost track recovery, candidate ambiguity, and object reappearance. Compared with simple feature concatenation, the cross-interaction approach enables the model to learn conditional relationships across modalities and improves the reliability of moving-object identification in crowded scenes.

Keywords: moving object identification; multimodal comparison; cross-interaction; object tracking; re-identification; crowded scenes.

1. Introduction

Moving object identification is an important task in video surveillance, smart monitoring systems, traffic analysis, and sports video analytics. The main goal is to determine whether two observations in video frames belong to the same moving object. This task is relatively simple when the object is clearly visible, and the scene contains only a few objects. However, in crowded video scenes, the identification problem becomes more difficult due to occlusion, similar appearances, object interactions, and temporary disappearances.

A common approach for object identification is to compare appearance features extracted from detected objects. Deep appearance descriptors are widely used in object tracking and person re-identification systems. For example, Deep SORT uses a deep appearance metric to improve identity association in multi-object tracking [1]. Person re-identification studies also show that learned visual embeddings can provide useful identity-related information [2], [3]. However, an appearance-only comparison is not always reliable. Two different people may wear similar clothes, while the same person may appear different due to pose, lighting, scale, or partial visibility.

Therefore, a more reliable identification framework should use several complementary modalities. In addition to appearance, geometric features can describe the scale and shape of the bounding box; spatial features can describe physical location and movement feasibility; contextual features can describe crowd density and ambiguity; reliability features can describe detection quality and visibility; and clothing-colour features can provide additional visual support.

However, simply concatenating all modalities is insufficient because it does not explicitly model their relationships. For example, appearance similarity should be trusted more when visibility is high; spatial distance should be more important when the time interval is short; colour similarity should be used carefully when visibility is low; and geometry may be less reliable under strong perspective distortion.

To solve this problem, this paper proposes a cross-interaction-based multimodal feature comparison method. The proposed method projects all modalities into a common latent space and then learns the interaction between feature groups. An MLP classifier estimates the final similarity score. This allows the model to decide whether two detections belong to the same identity by jointly analyzing multiple types of evidence.

2. Proposed Method

The proposed identification framework is based on multimodal feature comparison. The purpose of the method is to determine whether two observations correspond to the same moving object. In crowded video scenes, a single appearance descriptor may be unreliable because different objects may look similar, and the same object may be affected by pose, scale, visibility, occlusion, and illumination changes.

The proposed method consists of four main stages:

1. multimodal feature extraction;
2. projection of heterogeneous features into a common latent space;
3. cross-interaction-based pairwise feature comparison;
4. similarity estimation using an MLP classifier.

The general idea is that each detected object is represented by several complementary feature groups. Then, two candidate objects are compared using both similarity and dissimilarity information. Instead of using simple concatenation, the proposed method uses cross-interaction to learn how different modalities support or weaken each other.

3. Multimodal Object Representation

A moving object is represented using six modalities:

$$F = \{f^{app}, f^{geo}, f^{spa}, f^{ctx}, f^{rel}, f^{col}\} \quad (1)$$

where f^{app} is the appearance embedding, f^{geo} is the geometric feature vector, f^{spa} is the spatial feature vector, f^{ctx} is the contextual feature vector, f^{rel} is the reliability feature vector, and f^{col} is the clothing-colour feature vector.

3.1 Appearance Feature

The appearance feature is the main visual descriptor of the object. It can be extracted using a deep neural network or a re-identification model. In the proposed framework, the appearance embedding is represented as a 384-dimensional vector:

$$f^{app} \in R^{384} \quad (2)$$

This feature contains identity-related visual information such as clothing texture, body appearance, and general object appearance. However, appearance information may be unreliable under occlusion, illumination variation, or similar clothing.

3.2 Geometric Feature

The geometric feature vector contains bounding-box-related parameters:

$$f^{geo}=[w,h,A,r] \quad (3)$$

where w and h are the width and height of the bounding box, A is the area, and r is the aspect ratio. These features help evaluate the consistency of shape and scale across detections.

3.3 Spatial Feature

The spatial feature vector represents object location:

$$f^{spa}=[x_1,y_1,x_2,y_2,c_x,c_y,z] \quad (4)$$

where (x_1,y_1) and (x_2,y_2) are bounding-box coordinates, (c_x,c_y) are the centre coordinates, and z is the region identifier or normalized active-area coordinate. Spatial features help determine whether a candidate match is physically feasible.

3.4 Context Feature

The context feature vector describes the local environment around the object:

$$f^{ctx}=[d_{nn},n_r,\rho] \quad (5)$$

where d_{nn} is the nearest-neighbour distance, n_r is the number of nearby objects within radius r, and ρ is local density. Context is useful in crowded scenes because local density and nearby objects increase identity ambiguity.

3.5 Reliability Feature

The reliability vector can be defined as:

$$f^{rel}=[s,v,o,q] \quad (6)$$

where s is detector confidence, v is visibility, o is occlusion level, and q is quality score. This feature group helps the model estimate how much to trust the extracted features in the comparison.

3.5 Clothing-Colour Feature

The clothing-colour feature is obtained using HSV vectorization:

$$f^{col}=HSV(H,S,V) \quad (7)$$

HSV representation is useful because it separates hue, saturation, and value components. This makes clothing-colour comparison more stable than raw RGB comparison under moderate illumination variation. Colour is not sufficient as the main identity descriptor, but it can provide useful supporting information.

4. Modalities Used in Multimodal Comparison

The proposed multimodal comparison framework uses different feature groups to capture complementary information about each moving object. Each modality has a specific role in the identification process, ranging from identity-related visual representation to spatial feasibility, contextual ambiguity estimation, and reliability-aware decision-making. Table 1 summarises the modalities used in the proposed method and their roles in moving object identification.

Table 1. Modalities used in the proposed multimodal comparison method

Modality	Feature examples	Identification role

Appearance	384-dimensional embedding	Main identity-related visual descriptor
Geometry	Width, height, area, aspect ratio	Shape and scale consistency
Spatial	Centre, active region, normalized coordinates	Physical feasibility of matching
Context	Nearest-neighbour distance, local density	Ambiguity and crowd-risk estimation
Reliability	Confidence, visibility, occlusion, quality	Trustworthiness of comparison
Color	HSV clothing vector	Supporting clothing-based cue

The main advantage of using multiple modalities is that the system does not depend on a single feature type. In clean frames, appearance may dominate the matching decision. In crowded frames, context and reliability become more important. During reappearance after occlusion, spatial feasibility and temporal consistency become essential.

5. Projection into a Common Latent Space

Before fusion, all modalities are projected into a common latent space. This is necessary because the original features have different dimensions and different statistical properties.

For each modality mmm, the projection is defined as:

$$z_m = P_m(f_m) \quad (8)$$

where P_m is a projection function for modality mmm. The projected multimodal feature set is:

$$Z = \{z_{app}, z_{geo}, z_{spa}, z_{ctx}, z_{rel}, z_{col}\} \quad (9)$$

This projection allows heterogeneous modalities to be represented in the same latent dimension d.

Table 2. Feature dimensions before and after projection

Feature group	Original dimension	Projection output
Appearance	384	d
Geometry	4	d
Spatial	5 or grid-dependent	d
Context	3–4	d
Reliability	2–4	d
Color	Histogram-dependent	d

A simple concatenation of all projected modalities can be written as:

$$Z_{concat} = [z_{app} \parallel z_{geo} \parallel z_{spa} \parallel z_{ctx} \parallel z_{rel} \parallel z_{col}] \quad (10)$$

However, concatenation does not explicitly model relationships between modalities. For example, appearance similarity should be trusted more when reliability is high; spatial distance should be more important when the time interval is short; colour similarity should be used carefully when visibility is low; and geometry may be less reliable under strong perspective distortion. Therefore, the proposed method uses cross-interaction rather than simple concatenation.

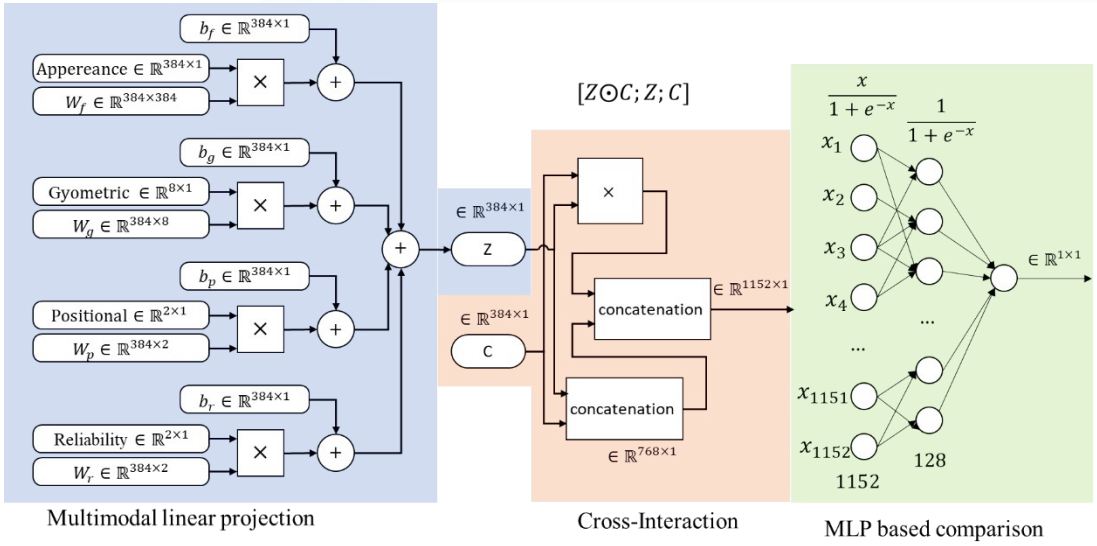

Fig. 1. Projection of heterogeneous modalities into a common latent space and similarity estimation using MLP.

6. Cross-Interaction-Based Feature Comparison

For two objects i and j, modality-wise comparison features are formed as:

$$c_m(i,j)=[z_m^i \odot z_m^j \parallel | z_m^i - z_m^j |] \quad (11)$$

where $\odot$ denotes element-wise multiplication. The element-wise product captures similarity, while the absolute difference captures dissimilarity.

The cross-interaction representation is defined as:

$$C(i,j)=\Phi(c_{app}, c_{geo}, c_{spa}, c_{ctx}, c_{rel}, c_{col}) \quad (12)$$

where Φ is the cross-interaction function. This function allows the model to learn relationships between modalities.

The cross-interaction function can learn conditions such as:

- Appearance similarity is more reliable when visibility is high.
- Spatial distance is more important when the frame interval is short.
- Clothing colour is useful when the full body is visible.
- Geometry is less reliable when the camera perspective changes.
- context becomes more important in crowded scenes.
- Reliability features can control how much the model trusts other modalities.

This is the main difference between cross-interaction and simple concatenation. Concatenation only combines features, whereas cross-interaction learns how they influence each other.

7. Similarity Estimation Using MLP

After cross-interaction, the final similarity probability is estimated using an MLP:

$$p_{ij}=\sigma(MLP(C(i,j))) \quad (13)$$

where σ is the sigmoid activation function, the output p_{ij} represents the probability that two detections belong to the same identity.

The binary decision rule is:

$$\hat{y}_{ij}=\begin{cases}1, & p_{ij}\geq\tau\\ 0, & p_{ij}<\tau\end{cases} \quad (14)$$

where τ is the decision threshold, if p_{ij} is greater than or equal to the threshold, the two detections are considered to belong to the same object. Otherwise, they are treated as different objects.

8. Training Objective

The Binary Cross-Entropy loss is used for pairwise matching:

$$L_{BCE} = -[y_{ij} \log(p_{ij}) + (1 - y_{ij}) \log(1 - p_{ij})] \quad (15)$$

where y_{ij} is the ground-truth label, suppose the two detections belong to the same identity, $y_{ij} = 1$ if they belong to different identities, $y_{ij} = 0$.

The training process encourages the model to assign higher similarity scores to positive pairs and lower scores to negative pairs. Hard negative pairs are especially important because they help the model distinguish visually similar but different objects.

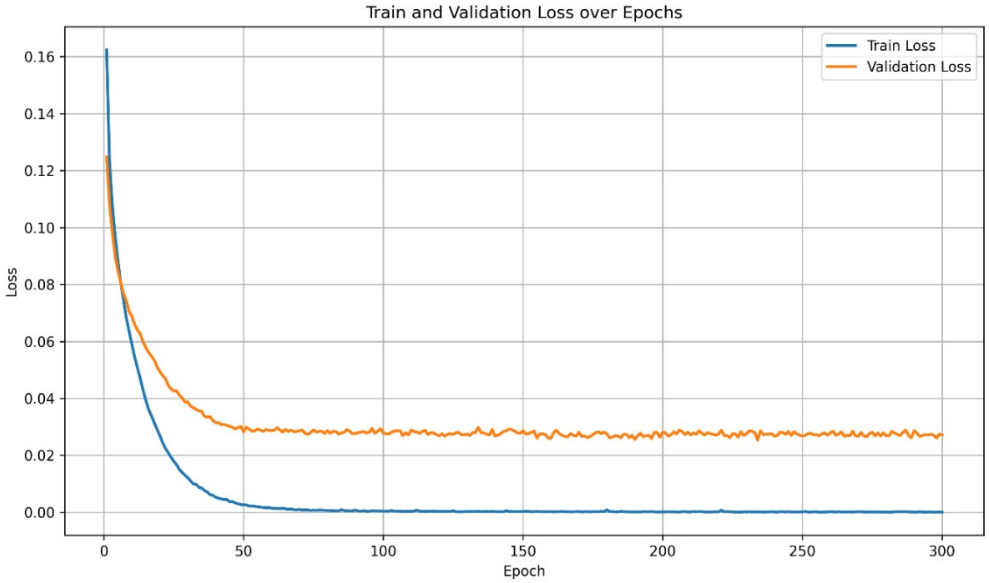

Fig. 2. Training results of the multimodal linear projection model.

Figure 2 illustrates the training behaviour of the projection module that transforms heterogeneous feature groups into a common latent space.

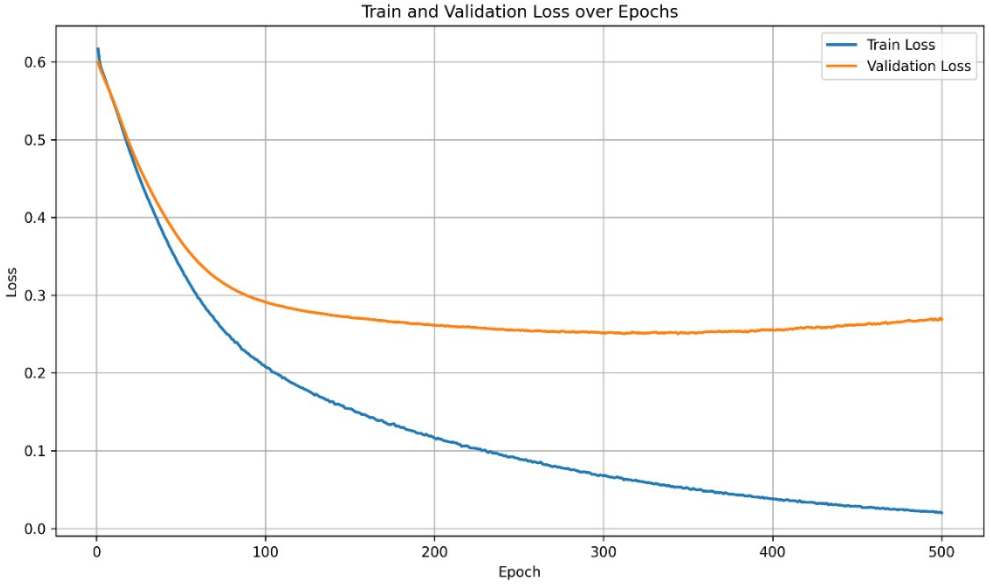

Fig. 3. Training results of the MLP-based comparison model.

Figure 3 shows the training performance of the MLP classifier used to estimate the similarity probability between two candidate detections after cross-interaction-based feature comparison.

9. Conclusion

This paper presented a cross-interaction-based multimodal feature comparison method for moving object identification in crowded video scenes. The proposed framework represents each object using appearance, geometry, spatial, context, reliability, and clothing-color features. These heterogeneous modalities are projected into a common latent space and compared using element-wise product and absolute difference operations.

The cross-interaction function learns relationships between modalities and allows the model to use different feature groups depending on scene conditions adaptively. The final similarity probability is estimated using an MLP classifier and converted into a binary decision using a threshold.

The proposed method is especially useful in difficult situations such as occlusion, lost track recovery, candidate ambiguity, and reappearance. Since the multimodal comparison module is mainly activated in difficult cases, the method strikes a balance between computational efficiency and identification reliability.

Adabiyotlar, References, Литературы:

1. N. Wojke, A. Bewley, and D. Paulus, “Simple online and real-time tracking with a deep association metric,” in Proceedings of the IEEE International Conference on Image Processing, pp. 3645–3649, 2017.
2. M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 2872–2893, 2022.
3. H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
4. X. Zheng, J. Zhu, Y. Sun, and Z. Zheng, "Multimodal person re-identification based on transformer relation regularisation," Information Fusion, vol. 104, article 102128, 2024.
K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, “Cross-modality transformer for visible-infrared person re-identification,” in Proceedings of the European Conference on Computer Vision, pp. 480–496, 2022.