УДК 004.932.4

ЕВОЛЮЦІЯ ПРОСТОРОВО-ЧАСОВИХ МЕТОДІВ ЗНЕШУМЛЕННЯ ЦИФРОВОГО ВІДЕО

Лисенко. І.А.
email: illia.lysenko3@nure.ua
Науковий керівник – канд. техн. наук, доц. Гусарова І.Г.
Харківський національний університет радіоелектроніки,
кафедра прикладної математики
м. Харків, Україна

The work reviews the evolution of spatio-temporal video denoising methods, from classical nonlocal approaches to modern neural network-based models. Particular attention is given to V-BM3D and V-BM4D models, including the mathematical description of the noisy video model, collaborative filtering, and weighted aggregation of estimates. The work further highlights the transition to deep learning methods that use temporal dependencies between neighboring frames for more effective video restoration.

Цифрове відео є важливим джерелом інформації в сучасних інформаційних системах, а його якість безпосередньо впливає на ефективність подальшої обробки системами комп'ютерного зору. Одним із головних чинників погіршення якості відео є шум – випадкове спотворення яскравості чи кольору пікселів, що ускладнює аналіз відеоданих. Тому знешумлення відео є важливою задачею комп'ютерного зору.

Одними з ключових етапів розвитку просторово-часових методів знешумлення відео стали V-BM3D і V-BM4D [1], що ґрунтуються на пошуку подібних фрагментів між кадрами та їх спільній фільтрації у трансформованій області. Нехай зашумлене відео описується моделлю

$$z(x,t) = y(x,t) + \eta(x,t), \tag{1}$$

де $y(x,t)$ – істинний відеосигнал, $z(x,t)$ – спостережуване зашумлене відео, $\eta(x,t)$ – адитивний білий гаусів шум, x – просторова координата, t – номер кадру. Задача полягає у побудові знешумленого наближення $\hat{y}$ до оригінального відеосигналу y зі спостереження z.

У методі V-BM3D для кожного опорного блока відбувається пошук подібних 2D-блоків в просторово-часовому околі, після чого вони об'єднуються у 3D-групу та спільно фільтруються. Алгоритм є двоетапним. На першому етапі формується базове наближення відео за допомогою групування і колаборативної фільтрації з жорстким порогуванням:

$$\hat{Y}_{S_x} = T_{3D}^{-1}(\mathrm{HARDTHR}(T_{3D}(Z_{S_x}),\ \lambda_{3D}\sigma)), \tag{2}$$

де $\hat{Y}_{S_x}$ – наближення групи блоків, Z_{S_x} – група подібних зашумлених блоків, T_{3D} – тривимірне перетворення, HARDTHR – оператор жорсткого порогування коефіцієнтів, σ – середньоквадратичне відхилення шуму, $\lambda_{3D}\sigma$ – поріг фільтрації. Базове наближення $\hat{y}^{basic}$ формується шляхом агрегації поблокових

наближень $\hat{Y}_{S_x}^{x}$ за допомогою зваженого усереднення. На другому етапі базове наближення використовується для уточненого групування, а жорстке порогування замінюється колаборативною фільтрацією Вінера:

$$\hat{Y}_{S_x} = T_{3D}^{-1}\left(T_{3D}(Z_{S_x}) \cdot \frac{[T_{3D}(\hat{Y}_{S_x}^{basic})]^2}{[T_{3D}(\hat{Y}_{S_x}^{basic})]^2 + \sigma^2} \right), \quad (3)$$

де $\hat{Y}_{S_x}^{basic}$ – базове наближення групи блоків, отримане на першому етапі. Як і на першому етапі, кінцеве наближення $\hat{y}^{final}$ формується шляхом агрегації поблокових наближень $\hat{Y}_{S_x}^{x}$ за допомогою зваженого усереднення.

Метод V-BM4D використовує подібні просторово-часові об'єми відео, тобто фрагменти, які охоплюють не лише локальну область кадру, а й її зміну в часі вздовж траєкторії руху. Для кожного опорного об'єму проводиться пошук подібних об'ємів, після чого вони об'єднуються у чотиривимірну групу. Колаборативна фільтрація для опорного об'єму з координатами (x_0, t_0) описується формулою

$$\hat{G}_y(x_0, t_0) = T_{4D}^{-1}(\mathrm{Y}(T_{4D}(G_z(x_0, t_0)))), \quad (4)$$

де $G_z(x_0, t_0)$ – 4D-група подібних просторово-часових об'ємів, T_{4D} – чотиривимірне лінійне перетворення, Y – оператор стискання коефіцієнтів.

Відфільтрована 4D-група складається з наближень окремих просторово-часових об'ємів:

$$\hat{G}_y(x_0, t_0) = \{\hat{V}_y(x_i, t_i) : (x_i, t_i) \in \mathrm{Ind}(x_0, t_0)\}, \quad (5)$$

де $\hat{V}_y(x_i, t_i)$ – наближення відповідного просторово-часового об'єму, а $\mathrm{Ind}(x_0, t_0)$ – множина індексів об'ємів, подібних до опорного.

Оскільки такі об'єми, як правило, перекриваються, для одних і тих самих координат відеосигналу виникає кілька оцінок. Тому остаточне наближення відео без шуму формується шляхом їх зваженої агрегації:

$$\hat{y} = \frac{\sum_{(x_0,t_0)\in X\times T}\left(\sum_{(x_i,t_i)\in \mathrm{Ind}(x_0,t_0)} w_{(x_0,t_0)} \cdot \hat{V}_y(x_i, t_i)\right)}{\sum_{(x_0,t_0)\in X\times T}\left(\sum_{(x_i,t_i)\in \mathrm{Ind}(x_0,t_0)} w_{(x_0,t_0)} \cdot \chi_{(x_i,t_i)}\right)}, \quad (6)$$

де $\hat{y}$ – наближення відео без шуму, $w_{(x_0,t_0)}$ – вага групи з опорним об'ємом (x_0, t_0), $\hat{V}_y(x_i, t_i)$ – наближення окремого просторово-часового об'єму, $\chi_{(x_i,t_i)}$ – характеристична функція області підтримки цього об'єму.

Перші нейромережеві підходи знешумлення відео виникли як поєднання нелокальних методів зі згортковими нейронними мережами. Однією з найвідоміших є модель VNLnet, де з груп фрагментів формується нелокальний вектор ознак, який прив'язується до відповідного пікселя поточного кадру. Далі згорткова нейронна мережа використовує ці ознаки для відновлення кадру без шуму. Подальший розвиток нейромережевих підходів привів до появи мультикадрових згорткових моделей, які використовують не лише

просторові ознаки окремих кадрів, а й часові залежності між ними. До таких моделей належать ViDeNN, DVDnet і FastDVDnet [2]. У ViDeNN мережа обробляє групу сусідніх кадрів і відновлює центральний кадр, орієнтуючись на задачу сліпого знешумлення. У DVDnet застосовується двоетапна схема: спочатку виконується просторове знешумлення окремих кадрів, а потім, після їх вирівнювання відносно центрального кадру, здійснюється часова агрегація інформації. У FastDVDnet від попереднього вирівнювання відмовилися: мережа безпосередньо відновлює центральний кадр із групи сусідніх, неявно враховуючи часові зв'язки. Це дало змогу спростити обробку, зменшити залежність від оцінки руху та підвищити швидкодію. Подальший розвиток методів знешумлення відео пов'язаний із появою рекурентних моделей, як, наприклад, EMVD [3], в яких інформація з попередніх кадрів не щоразу обробляється заново, а передається далі у вигляді внутрішнього представлення, яке й використовується для відновлення кадру. Це дає змогу накопичувати часовий контекст вздовж усієї відеопослідовності та ефективніше використовувати часові залежності між кадрами. Це знижує вартість обчислень, проте для таких підходів характерна чутливість до накопичення помилок. Сучасні трансформерні архітектури активно застосовуються в задачах знешумлення відео, використовуючи механізм уваги для встановлення відповідностей між фрагментами сусідніх кадрів, як у моделі VRT [4]. У моделі RVRT [4] цей підхід був розвинений завдяки рекурентній схемі обробки та механізму деформованої уваги, що дає змогу точніше враховувати рух і підвищувати якість відновлення відео. Розвиток просторово-часових методів знешумлення відео відображає перехід від нелокальних підходів та алгоритмів фільтрації до сучасних нейромережевих моделей, здатних ефективніше використовувати часові залежності між кадрами та забезпечувати вищу якість відновлення відеоданих.

Список використаних джерел:

1. Maggioni M., Boracchi G., Foi A., Egiazarian K. Video denoising using separable 4D nonlocal spatiotemporal transforms. *Proceedings of SPIE. Image Processing: Algorithms and Systems IX*. 2011. Vol. 7870. Art. 787003.

2. Tassano M., Delon J., Veit T. DVDnet: A Fast Network for Deep Video Denoising. *2019 IEEE International Conference on Image Processing (ICIP)*. 2019. P. 1805–1809.

3. Maggioni M., Huang Y., Li C., Xiao S., Fu Z., Song F. Efficient Multi-Stage Video Denoising with Recurrent Spatio-Temporal Fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. P. 3467–3476.

4. Liang J., Fan Y., Xiang X., Ranjan R., Ilg E., Green S., Cao J., Zhang K., Timofte R., Van Gool L. Recurrent Video Restoration Transformer with Guided Deformable Attention. *Advances in Neural Information Processing Systems*. 2022. Vol. 35. P. 378–393.